

# Long-Range Indoor Navigation With PRM-RL

Anthony Francis <sup>✉</sup>, Member, IEEE, Aleksandra Faust <sup>✉</sup>, Senior Member, IEEE, Hao-Tien (Lewis) Chiang <sup>✉</sup>, Member, IEEE, Jasmine Hsu <sup>✉</sup>, Member, IEEE, J. Chase Kew <sup>✉</sup>, Member, IEEE, Marek Fiser, Member, IEEE, and Tsang-Wei Edward Lee <sup>✉</sup>

**Abstract**—Long-range indoor navigation requires guiding robots with noisy sensors and controls through cluttered environments along paths that span a variety of buildings. We achieve this with PRM-RL, a hierarchical robot navigation method in which reinforcement learning (RL) agents that map noisy sensors to robot controls learn to solve short-range obstacle avoidance tasks, and then sampling-based planners map where these agents can reliably navigate in simulation; these roadmaps and agents are then deployed on robots, guiding them along the shortest path where the agents are likely to succeed. In this article, we use probabilistic roadmaps (PRMs) as the sampling-based planner, and AutoRL as the RL method in the indoor navigation context. We evaluate the method with a simulation for kinematic differential drive and kinodynamic car-like robots in several environments, and on differential-drive robots at three physical sites. Our results show that PRM-RL with AutoRL is more successful than several baselines, is robust to noise, and can guide robots over hundreds of meters in the face of noise and obstacles in both simulation and on robots, including over 5.8 km of physical robot navigation.

**Index Terms**—Probabilistic roadmaps (PRMs), reinforcement learning (RL), robotics, navigation, sampling-based planning.

## I. INTRODUCTION

LONG-RANGE indoor robot navigation requires human-scale robots, as shown in Fig. 1(a), to move safely over building-scale distances, as shown in Fig. 1(b). To robustly navigate long distances in novel environments, we factor the problem into long-range path planning and end-to-end local control, while assuming the robot has mapping and localization. Long-range path planning finds collision-free paths to distant goals not reachable by local control [43]. End-to-end local control produces feasible controls to follow ideal paths while avoiding obstacles, e.g., [40] and [24], and compensating for noisy sensors and localization [12]. We enable end-to-end local control to inform long-range path planning through sampling-based planning.

Sampling-based planners, such as probabilistic roadmaps (PRMs) [39] and rapidly exploring random trees (RRTs) [42],

Manuscript received March 4, 2019; accepted January 16, 2020. Date of publication April 15, 2020; date of current version August 5, 2020. This research was funded solely by Google. This article was recommended for publication by Associate Editor F. Stulp and Editor Francois Chaumette upon evaluation of the reviewers' comments. (Corresponding author: Anthony Francis.)

The authors are with the Robotics at Google, Mountain View, CA 94043 USA (e-mail: centaur@google.com; sandrafaust@google.com; lewispro@google.com; hellojas@google.com; jkew@google.com; mfiser@google.com; tsangwei@google.com).

Color versions of one or more of the figures in this article are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TRO.2020.2975428



Fig. 1. Long-range indoor navigation task. (a) Approximately cylindrical differential drive robot. (b) Deployment environments are office buildings.

[44], plan efficiently by approximating the topology of the configuration space ( $\mathcal{C}$ ), the set of all possible robot poses, with a graph or tree constructed by sampling points in the collision-free subset of configuration space ( $\mathcal{C}_{\text{free}}$ ), and connecting these points if there is a collision-free local path between them. Typically, these local paths are created by line-of-sight tests or an inexpensive local planner, and are then connected in a sequence to form the full collision-free path.

Regardless of how a planner generates a path, executing a path requires handling sensor noise, unmodeled dynamics, and environment changes. Recently, reinforcement learning (RL) agents [41] have solved complex robot control problems [68], generated trajectories under task constraints [22], demonstrated robustness to noise [21], and learned complex skills [57], [55], making them good choices to deal with task constraints. Many simple navigation tasks only require low-dimensional sensors and controls, such as lidar and differential drive, and can be solved with easily trainable networks [7], [29], [71]. However, as we increase the problem's complexity by requiring longer episodes or providing only sparse rewards [20], RL agents become harder to train and do not consistently transfer well to new environments [35], [34].

Long-range navigation presents all these challenges. Sparse rewards and long episodes make long-range agents hard to train. On complex maps, short-range agents are vulnerable to local minima such as wide barriers and narrow passages. Even within deployment categories, environments have vast variety: the SunCG dataset had 45 000 houselike environments [64], and the US alone has over 5.6 million office buildings [9].

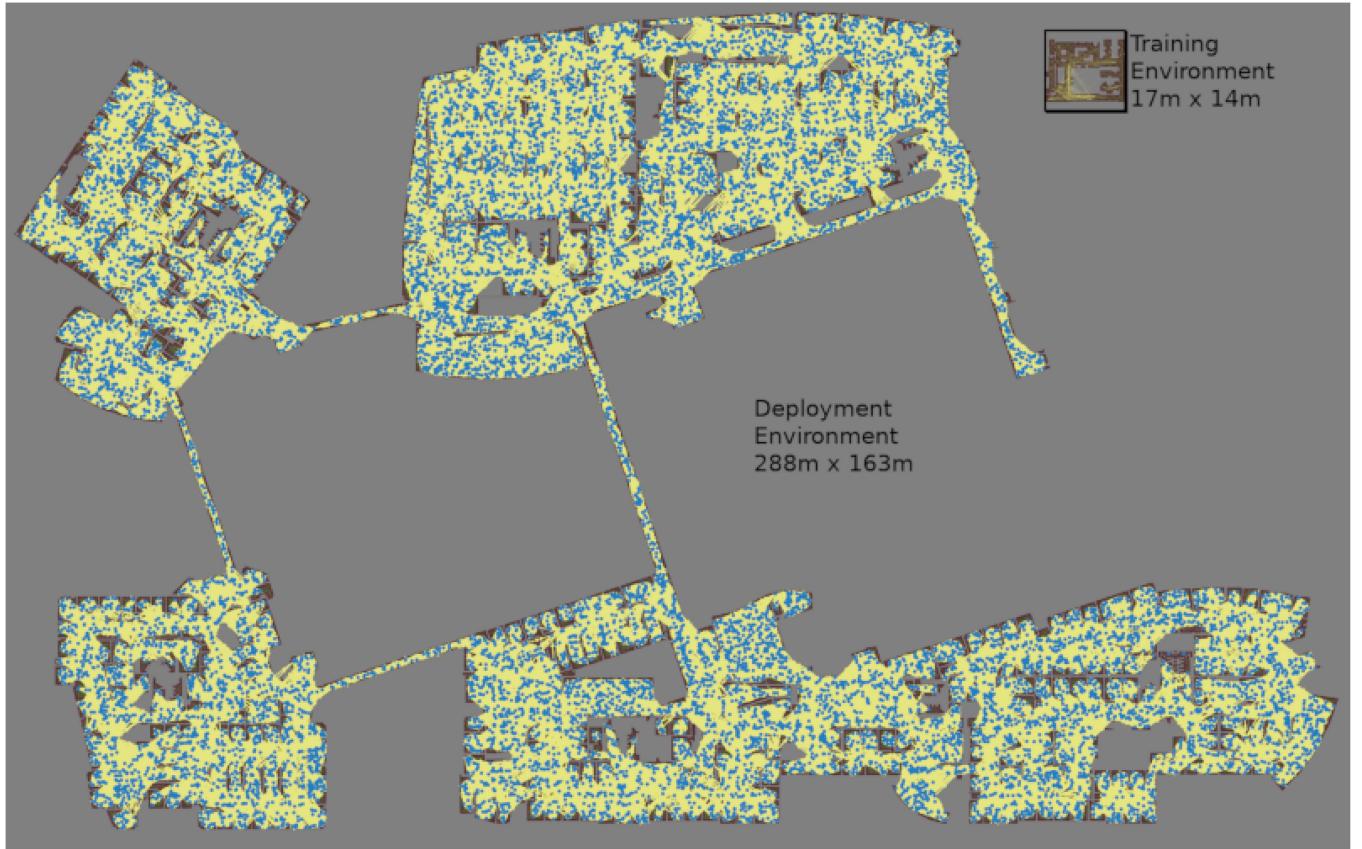


Fig. 2. Quad-building complex— $288 \times 163$  m: A large roadmap derived from real building plans, which PRM-RL successfully navigated 57.3% of the time in simulation. The connected segment in the upper center corresponds to the upper floor of Building 1 used in our evaluations and contains the space where we collected our SLAM map in Fig. 6. Blue dots are sampled points and yellow lines are roadmap connections navigable in simulation with the AutoRL policy. This roadmap has 15 900 samples and had 1.4 million candidate edges prior to connection attempts, of which 689 000 edges were added. It took 4 days to build using 300 workers in a cluster, requiring 1.1 billion collision checks. The upper right inset is the training environment from Fig. 3(a), to scale; the quad-building complex is approximately 200 times larger in map area.

We present PRM-RL, an approach to long-range navigation, which combines PRMs and RL to overcome each other's shortfalls. In PRM-RL, an RL agent learns a local point-to-point (P2P) task, incorporating system dynamics and sensor noise independent of long-range environment structure. The agent's learned behavior then influences roadmap construction; PRM-RL builds a roadmap by connecting two workspace points only if the agent consistently navigates between them in configuration space without collision, thereby learning the long-range environment structure. PRM-RL roadmaps perform better than roadmaps based on pure  $\mathcal{C}_{\text{free}}$  connectivity because they respect robot dynamics. RL agents perform better with roadmap guidance, avoiding local minima. PRM-RL, thus, combines PRM efficiency with RL resilience, creating a long-range navigation planner that not only avoids local-minima traps, but transfers well to new environments, as shown in Fig. 2, where a policy trained on a small training environment scales to a quad-building complex almost 200 times larger in map area.

This article contributes PRM-RL as a hierarchical kinodynamic planner for navigation in large environments for robots with noisy sensors. This article is a journal extension of our conference paper [23], which contributes the original PRM-RL method. Here, we investigate PRM-RL in the navigation context and make the following contributions beyond the original paper:

1) Algorithm 2 for PRM-RL roadmap building; 2) Algorithm 3 for robot deployment; 3) PRM-RL application to kinodynamic planning on a car model with inertia; and 4) in-depth analysis of PRM-RL, including: 4.1) correlation between the quality of the local planner and the overall hierarchical planner; 4.2) impact of improving planners and changing parameters on PRM-RL computational time complexity; 4.3) impact of a robust local planner on the effective connectivity of samples in the graph; and 4.4) resilience of PRM-RL to noise and dynamic obstacles. All the evaluations and experiments are new and original to this article. We evaluate PRM-RL with a more effective local planner [12], compare it in simulation against six baselines in eight different buildings, and deploy it to three physical robot testbed environments.

Overall, we show improved performance over both baselines and our prior work, more successful roadmaps, and easier transfer from simulation to robots, including a 37.5% increase in navigation success over [23], while maintaining good performance despite increasing noise. We also show that only adding edges when agents can always navigate them makes roadmaps cheaper to build and improves navigation success; denser roadmaps also have higher simulated success rates, but at substantial roadmap construction cost. Floorplans are not always available or up to date, but we show roadmaps built from SLAM maps close the



Fig. 3. Environments used for indoor navigation are derived from real building plans. (a) Smallest environment is used to train the RL agent for faster training and iteration. (b)–(d) PRMs for deployment environments are built using agents trained in the training environment. Red regions are deemed too close to obstacles and cause episode termination when the robot enters them; light gray is free space from which the starts and goals are selected. Blue lines connect PRM waypoints, and the RL agent's executed trajectory is black.

simulation to reality gap by producing planners, which perform almost as well on robot as they do in simulation. SLAM-based PRM-RL enables real robot deployments with up to +200 m collision-free trajectories at three different sites on two different robots with success rates as high as 92.0%. We also show that PRM-RL functions well on robots with dynamic constraints, with an 83.4% success rate in simulation.

While this article focuses on navigation, the analysis and empirical findings will be of interest to the wider motion planning community for two reasons. First, PRM-RL presented here is an example of a hierarchical motion planner that factors models of sensor, localization, and control uncertainties into roadmap construction, resulting in planners that perform as well in simulation as they do on robots. Second, we present a comprehensive analysis of the tradeoffs between performance and computational complexity and the interplay between local and global planners that is not specific to navigation.

## II. RELATED WORK

### A. Probabilistic Roadmaps

PRMs [39] have been used in a wide variety of planning problems from robotics [30], [51] to molecular folding [3], [59], [66]. They have also been integrated with RL for state-space

dimensionality reduction [49], [59] by using PRM nodes as the state space for the RL agent. In contrast, our work applies RL to the full state space as a local planner for PRMs. In prior work for an aerial cargo delivery task, we trained RL agents to track paths generated from PRMs constructed using a straight-line local planner [22]. Researchers have modified PRMs to handle moving obstacles [32], [61], noisy sensors [50], and localization errors [2], [4]. Safety PRM [50] uses probabilistic collision checking with a straight-line planner, associating a measure of potential collision with all nodes and edges. All those methods address one source of errors at a time. In contrast, PRM-RL uses an RL-based local planner capable of avoiding obstacles and handling noisy sensors and dynamics, and at the node connection phase, the RL local planner does Monte Carlo path rollouts with deterministic collision checking. We only add edges if the path can be consistently navigated within a tunable success threshold. Additionally, PRMs built with straight-line geometric planners require a tracking method (often model-based, such as [67]) to compensate for sensor and dynamics uncertainties not known at the planning time. PRM-RL eliminates the need for path tracking and mathematical models for sensor and dynamics uncertainties, because sensor-to-control RL local planners are model-free, and the Monte Carlo rollouts ensure that the feasibility is accounted when connecting the nodes.

PRMs are easy to parallelize, either through parallel edge connections [5], sampling [8], or building subregional roadmaps [17] in parallel. To speed up building large-scale roadmaps, we use an approach similar to [5] across different computers in a cluster. Individual Monte Carlo rollouts to connect edges can be parallelized across multiple processors or run sequentially to allow for early termination.

1) *RL in Motion Planning*: RL has recently gained popularity in solving motion planning problems for systems with unknown dynamics [41], and has enabled robots to learn tasks that have been previously difficult or impossible [1], [10], [45]. For example, deep deterministic policy gradient (DDPG) [46] works with high-dimensional continuous state/action spaces and can learn to control robots using unprocessed sensor observations [45].

Researchers have successfully applied deep RL to navigation for robots, including visual navigation with simplified navigation controllers [7], [16], [29], [56], [62], [72], more realistic controllers in game-like environments [6], [15], [54], and extracting navigation features from realistic environments [10], [26]. In local planner settings similar to ours, differential drive robots with 2-D lidar sensing, several approaches emerged recently using asynchronous DDPG [65], expert demonstrations [60], DDPG [47], curriculum learning [70], and AutoRL [12]. While any sensor-to-controls obstacle-avoidance agent could be used as both a local planner in PRM-RL and a controller for reactive obstacle avoidance, we choose AutoRL for its simplicity of training, as it automates the search for reward and network architecture.

Recent works [48], [70] learn planning in 2-D navigation mazes and measure transferability of learning to new environments. Using the terminology of this article, this work falls between long-range planning and local control. Their navigation environments vary between task instances, as in both our AutoRL and PRM-RL setting. The mazes appear to be similar in size and complexity to the maps used in our local planning [12], [23]. However, those methods, based on Q-learning, use discretized actions to produce their paths, leading to three consequences. First, the robot's action discretization inherently limits how close to the optimal path the solution can reach. Second, the method must have a lower-level controller or steering function that tracks the resulting path. Third, while these planners avoid obstacles, they do not take robot dynamics or path feasibility into account. In our work, we rely on continuous action RL algorithms, which can better approximate optimal paths than discrete actions, eliminate the tracking controller, and learn dynamically feasible steering functions that produce linear and angular velocities [13].

### B. Hierarchical Planners With RL

Several recent works feature hierarchical planners combined with RL, either over a grid [19] or manually selected waypoints [37]. These works connect roadmap points with a straight-line planner and use the RL agent as an execution policy at run-time. We use the obstacle-avoidance RL policy as both an execution policy and a local planner for connecting the edges

in the roadmap. This approach results in roadmap connectivity that is tuned to the abilities of the particular robot.

## III. PROBLEM STATEMENT

This section defines key terms for path planning, for trajectory planning, for the navigation problem for robots with noisy depth sensors and actuators, and for the differential drive and car-like robots, which are our primary focus in this article.

The *configuration space*,  $\mathcal{C}$ , is the set of possible robot *poses*. The configuration spaces for differential drive and car-like robots are  $\mathcal{C}_{dd} = \mathbb{R}^2 \times S^1$  and  $\mathcal{C}_{car} = \mathbb{R}^2 \times S^1 \times S^1$ , respectively. The *workspace*,  $\mathcal{W}$ , is the physical space that a robot operates in with dimensionality  $D_{\mathcal{W}}$  of 2 or 3. The workspace and the robot's kinematics divide the configuration space into valid ( $\mathcal{C}_{free}$ ) and invalid partitions.  $\mathcal{C}_{free}$  is a set of all robot poses that are free of self-collision, collisions in the workspace, and satisfy relevant kinodynamic constraints. To that end, we consider the workspace to be a closed segment on a 2-D manifold, and model the robots' kinematics with a unicycle or Type (2,0) model [63] for differential drive robots, and single-track model [58] for car-like robots.

The robot *state space* is the full internal state of the robot including pose, velocity, observations, etc. We assume this state to be hidden and nonobservable. The observable state space,  $O \subset \mathbb{R}^{N_s \theta_n} \times \mathcal{C}_{free}^2$ , is the same for both robot types and consists of sensor observations ( $N_s$  lidar rays observed over the last  $\theta_n$  discrete time steps) as well as the current and goal robot poses, assumed to be in  $\mathcal{C}_{free}$ . The robot *action space*,  $A \subset \mathbb{R}^2$ , consists of linear and angular velocities  $\mathbf{a} = (v, \phi) \in A$  for both types of robots. We assume sensor observation and actuators to be noisy. The sensor observations are produced by a sensor process  $F_s : \mathcal{C}_{free} \rightarrow O$  that can be modeled as a combination of inherent sensor dynamics and a source of noise:  $F_s(\mathbf{x}) \sim D_s(\mathbf{x}) + \mathcal{N}_s$ . Similarly, actions in the robot's action space  $A$  have a state-dependent effect  $F_a : \mathcal{C}_{free} \times A \rightarrow \mathcal{C}$ , which also can be modeled as a combination of inherent robot dynamics and a source of noise:  $F_a(\mathbf{x}, \mathbf{a}) \sim D_a(\mathbf{x}, \mathbf{a}) + \mathcal{N}_a$ .

A *path*,  $\mathcal{P}$ , is a sequence of workspace points  $\mathbf{p}_i \in \mathcal{W}, i \in [0, N]$  from the beginning  $p_0$  to the end  $p_N$  of the trajectory. A *valid path* consists of only valid waypoints:  $\forall \mathbf{p}_i \in \mathcal{P} : \mathbf{p}_i \in \mathcal{W} \cap \mathcal{C}_{free}, i = 1, \dots, N_p$ . We consider a *trajectory*,  $\mathcal{T}$ , to be a sequence of robot valid poses  $\mathbf{x}_j \in \mathcal{C}_{free}, j = 1, \dots, N_t$  such that  $\mathbf{x}_{j+1}$  is reachable from  $\mathbf{x}_j$  within  $\mathcal{C}_{free}$  under the robot's kinematic model within a fixed discrete time step  $\Delta T$ , for any two consecutive points  $\mathbf{x}_{j+1}, \mathbf{x}_j, j = 1, \dots, N_t - 1$ . We assume that the robot is at rest at the beginning and end of the trajectory, i.e.,  $\dot{\mathbf{x}}_1 = \mathbf{0}, \dot{\mathbf{x}}_N = \mathbf{0}$ .

In this article, a *P2P* policy,  $\pi : O \rightarrow A$ , maps robot observations to linear and angular velocities in order to generate trajectories. Given a valid start configuration  $\mathbf{x}_S$  and a policy  $\pi$ , an *executed trajectory*  $\mathcal{T}$  is a sequence of configuration states that result from drawing actions from the policy and its noise processes:  $\mathcal{T} : \mathbf{x}_0 = \mathbf{x}_S \wedge \mathbf{x}_i \sim F_a(\mathbf{x}_{i-1}, \pi(F_s(\mathbf{x}_{i-1})))$ . An executed trajectory is a *failure* if it produces a point that exits  $\mathcal{C}_{free}$ , or if it exceeds a task-specific time-step limit  $\mathcal{K}_\omega$  without

reaching a goal. Given a valid observable goal pose,  $x_G$ , a non-failed trajectory is a *success* if it reaches a point  $x_i$  sufficiently close to the goal with respect to a task-dependent threshold  $\|x_i - x_G\| \leq d_G$ , at which point the task has *completed*.

Graph search over PRMs creates a path,  $\mathcal{P}$ . Waypoints in the path serve as intermediate goals for the trajectory generating policy  $\pi$ . A *path following policy* with respect to a path  $\mathcal{P}$  and P2P policy  $\pi$ ,  $\pi_{pf}(x|\mathcal{P}, \pi)$ , produces a trajectory that traverses the waypoints in path  $\mathcal{P}$  using the P2P policy  $\pi$ . A *valid executable path* with respect to a P2P policy is a path, which the P2P policy can reliably execute to achieve task completion—guiding the agent from the start state  $x_S$  of  $\mathcal{P}$  to within  $d_G$  of the goal state  $x_G$  within  $\mathcal{K}_\omega$  time steps. Because noise makes execution stochastic, we define a path to be *reliable* if the policy’s probability of task completion using the path exceeds a task-dependent *success threshold*  $p_s$ .

The key problem that PRM-RL addresses is how to construct a *reliable path* and *path following policy* for a given P2P policy in the context of indoor navigation. To that end, we define an agent that performs its task without knowledge of the workspace topology as one in which the transfer function  $\dot{x} = f(x, a)$  that leads the system to task completion is only conditioned on what the robot can observe and what it has been commanded to do. Formally, we learn policies to control an agent that we model as a partially observable Markov decision process represented as a tuple  $(O, A, D, N, R, \gamma)$  of observations, actions, dynamics, noise, reward, and discount. The characteristics of the robot determine observations, actions, dynamics, and noise; these are continuous and multidimensional. Reward and discount are determined by the requirements of the task:  $\gamma \in (0, 1)$  is a scalar discount factor, whereas the reward  $R$  has a more complicated structure  $(G, r)$ , including a true scalar objective  $G$  representing the task and a weighted dense scalar reward  $r : O \times \mathbb{R}^{N_\theta} \rightarrow \mathbb{R}$ , based on observable features  $O$  and a reward parameterization  $\theta \in \mathbb{R}^{N_\theta}$ . We assume a presence of a simplified black-box simulator without knowing the full nonlinear system dynamics. The dynamics  $D$  and noise  $N$  are implicit in the real world but are encoded separately in simulation in the system dynamics and an added noise model.

#### IV. METHODS

The PRM-RL method has three stages: training an environment-independent local planner policy with RL, creating a roadmap specific to that local planner and an environment, and deploying that roadmap and policy to the environment for querying, trajectory generation, and execution. Fig. 4 illustrates the method.

First, in the training stage, as shown in Fig. 4(a), to enable a robot to perform a specific task, we train an agent with RL. For indoor navigation, that task is short-range P2P navigation end-to-end from sensors to actions. This task is independent of the environment in which the robot will eventually operate. The RL agent learns to perform a task on an environment comparable to the deployment environment, but smaller in size to make simulation faster and training more tractable. This is a Monte Carlo simulation process: we train multiple policies and select

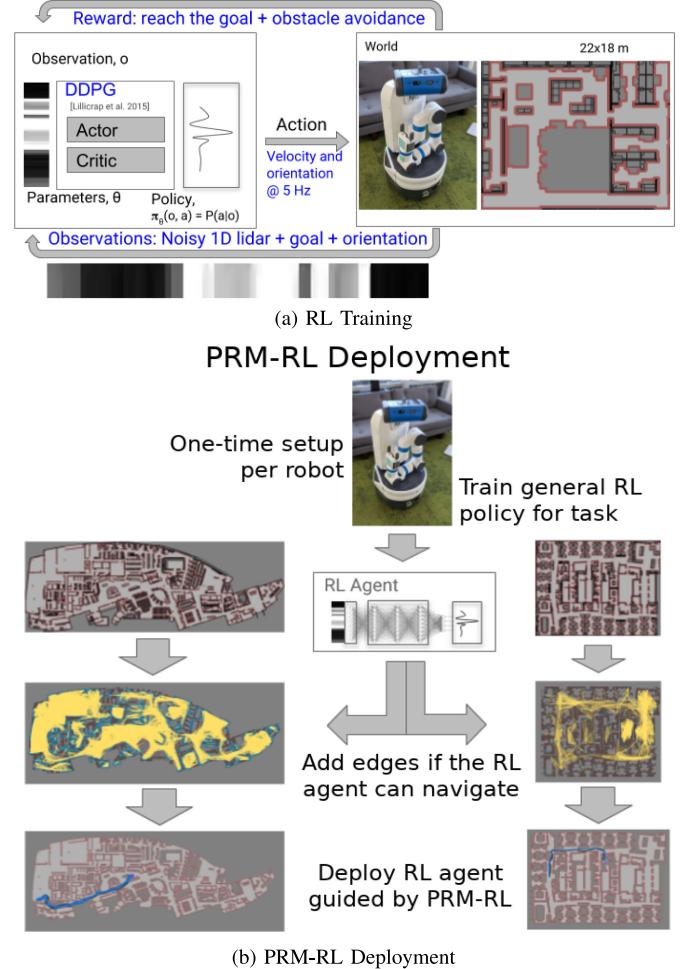


Fig. 4. PRM-RL approach. (a) RL learns a model of task and system dynamics. This enables the construction of a local planner and the generation of a PRM-RL roadmap. This roadmap and policy can then be executed on the robot using the same local planner. (b) Same policy can generate roadmaps for different floorplans, enabling deployment to many sites.

the fittest one for the next stage of PRM-RL, regardless of the learning algorithm used.

Next, in the creation phase, as shown in Fig. 4(b) upper, to prepare a robot to work in a specific environment, we use this best policy as a local planner to build a PRM with respect to the target site. Obstacle maps, such as floor plans or SLAM maps, can be used for any robot we wish to deploy as long as the policy transfers well to the real world. This is a one-time setup per robot and environment. Unlike PRM-SL, in which points are sampled from  $\mathcal{C}_{free}$ , PRM-RL samples points from the workspace  $\mathcal{W}$  but then rolls out trajectories in  $\mathcal{C}$ , adding an edge to the roadmap only when the agent can navigate it in simulation with greater probability than  $p_s$  over  $n_\omega$  trials. Rather than being determined by the geometry of free space, the resulting roadmap is tuned to capabilities of the particular robot, so different robots over the same floor plan may generate different roadmaps with different connectivity.

Finally, in the deployment phase, as shown in Fig. 4(b) lower, to perform the task in the environment, the constructed roadmap is queried to generate trajectories, which are executed

by the same RL agent used to generate the roadmap. In the geometric PRM-SL framework, querying a roadmap involves connecting the start and goal to the roadmap graph and finding the shortest path between them. In the simulation-based PRM-RL framework, we can optionally record additional data about executed trajectories to enable other trajectory metrics (such as minimal energy consumption, shortest time, and so on) that are not generally available with geometry-only approaches. At execution time, the RL agent navigates to the first waypoint. Once the agent is within  $d_G$  distance from the waypoint, the next waypoint becomes its new goal; the process repeats until the agent has traversed the whole trajectory.

### A. RL Agent Training

PRM-RL's global planner is strongly decoupled from the details of the local planner's construction and training. We explore this with two different agent models: differential drive and car-like robots.

1) *P2P for Differential Drive Robots*: The true objective of the P2P agent is to maximize the probability of reaching the goal without collisions

$$G_{\text{P2P}}(\mathbf{x}_i, \mathbf{x}_G) = \mathbb{I}(\|\mathbf{x}_i - \mathbf{x}_G\| < d_G) \quad (1)$$

where  $\mathbb{I}$  is an indicator function,  $\mathbf{x}_G$  is the goal position, and  $d_G$  is the goal radius. The zero-collision property is enforced by terminating episodes on collisions. The goal observation  $\mathbf{o}_g$  is the relative goal position in polar coordinates, which is readily available from localization. The reward for P2P for differential drive robots is the dot product of the parameters and reward components

$$R_{\theta_{r_{\text{DD}}}} = \theta_{r_{\text{DD}}}^\top [r_{\text{goal}} \ r_{\text{goalDist}} \ r_{\text{collision}} \ r_{\text{step}} \ r_{\text{turning}}]^\top \quad (2)$$

where  $r_{\text{goal}}$  is 1 when the agent reaches the goal and 0 otherwise,  $r_{\text{goalDist}}$  is the negative Euclidean distance to the goal,  $r_{\text{collision}}$  is 1 when the agent collides with obstacles and 0 otherwise,  $r_{\text{clearance}}$  is the distance to the closest obstacle,  $r_{\text{step}}$  is a constant penalty step with value 1, and  $r_{\text{turning}}$  is the negative angular speed. We train this model with AutoRL [12] over the DDPG [46] algorithm, which simultaneously finds the reward weights  $\theta_{r_{\text{DD}}}$  and trains the agent. AutoRL automates hyperparameter search in RL using an evolutionary approach. AutoRL takes as inputs a true objective used to evaluate the agent, a parameterized dense reward that the agent uses to train itself, and optionally neural network architecture and algorithm hyperparameters. To train the agent, AutoRL typically optimizes these hyperparameters in phases. First, given an arbitrary or hand-tuned architecture, it trains several populations of RL agents with different reward parameters and optimizes over the true objective. Optionally, a second phase repeats the process with the dense reward fixed while searching over neural network architectures instead.

2) *P2P for Car-Like Robots*: The true objective of P2P does not change for car drive, but because the turning radius of the car is limited and the car must perform more complex maneuvers, we choose a slightly different reward model

$$R_{\theta_{r_{\text{CM}}}} = \theta_{r_{\text{CM}}}^\top [r_{\text{goal}} \ r_{\text{goalProg}} \ r_{\text{collision}} \ r_{\text{step}} \ r_{\text{backward}}]^\top \quad (3)$$

where all values are the same as for diff drive except  $r_{\text{goalProg}}$  rewards the delta change of Euclidean distance to goal, and  $r_{\text{backward}}$  is the negative of backwards speed and 0 for forward speed. We dropped  $r_{\text{goalDist}}$ ,  $r_{\text{clearance}}$ , and  $r_{\text{turning}}$  to reduce the space of hyperparameter optimization based on analysis of which parameters seemed to have the most positive impact upon learning differential drive models. We train this model with hyperparameter tuning with Vizier [28] over the DDPG [46] algorithm in a different training regime in which the car model is allowed to collide up to ten times in training, but is still evaluated on the true objective of zero collisions.

### B. PRM Construction

The basic PRM method works by sampling robot configurations in the robot's configuration space, retaining only collision-free samples as nodes in the roadmap. PRMs then attempt to connect the samples to their nearest neighbors using a local planner. If an obstacle-free path between nodes exists, PRMs add an edge to the roadmap.

We modify the basic PRM by changing the way nodes are connected. Formally, we represent PRMs with graphs modeled as a tuple  $(V, E)$  of nodes and edges. Nodes are always in free space,  $V_i \in \mathcal{C}_{\text{free}}$ , and edges always connect valid nodes  $(V_i, V_j)$ , but we do not require that the line of sight  $\overline{V_i V_j}$  between those nodes is in  $\mathcal{C}_{\text{free}}$ , allowing edges that "go around" corners. Since we are primarily interested in robustness to noise and adherence to the task, we only connect configurations if the RL agent can consistently perform the P2P task between two points.

Algorithm 1 describes how PRM-RL adds edges to the PRMs. We sample multiple points from the configuration space, around the start and goal in the workspace, and attempt to connect the two points over  $n_\omega$  trials. An attempt is successful only if the agent reaches sufficiently close to the goal point. Even with a high success threshold, PRM-RL trajectories are not guaranteed to be collision-free because of sensor and action noise. To compute the total length of a trajectory, we sum the distances for all steps plus the remaining distance to the goal. The length we associate with the edge is the average of the distance of successful edges. The algorithm adds the edge to the roadmap if the success probability  $\text{success}_{\text{rate}}$  is above the threshold  $p_s$ .

The worst-case number of collision checks in Algorithm 1 is  $O(\mathcal{K}_\omega * n_\omega)$ , because multiple attempts are required for each edge to determine  $\text{success}_{\text{rate}}$ . Each trial of checking the trajectory can be parallelized with  $n_\omega$  processors; alternately, trajectory checking within Algorithm 1 can be serialized, terminating early if the tests fail too many times. Mathematically, for a given success threshold and desired number of attempts, at least  $n_s = \lceil p_s * n_\omega \rceil$  trials must succeed; therefore, we can terminate when  $n_s > p_s * n_\omega$ , or when the failures exceed the complementary probability  $n_f > (1 - p_s) * n_\omega$ . This can provide substantial savings if  $p_s$  is high, as shown in Section V-D. Much of PRM-RL construction can be parallelized; parallel calls to Algorithm 1 can speed roadmap construction, and roadmaps can be constructed in parallel.

We use a custom kinematic three-dimensional simulator, which provides drive models for agents and supports visual sensors such as cameras and lidars. The simulator also provides

**Algorithm 1;** PRM-RL AddEdge.

---

**Input:**  $s, g \in \mathcal{W} \cap \mathcal{C}_{\text{free}}$ : Start and goal in workspace.  
**Input:**  $p_s \in [0, 1]$  Success threshold.  
**Input:**  $n_\omega$ : Number of attempts.  
**Input:**  $d_G$ : Sufficient distance to the goal.  
**Input:**  $\mathcal{K}_\omega$ : Maximum steps for trajectory.  
**Input:**  $L(s)$ : Task predicate.  
**Input:**  $\pi$ : RL agent's policy.  
**Input:**  $D$  Generative model of system dynamics.  
**Output:**  $\text{add}_{\text{edge}}, \text{success}_{\text{rate}}, \text{length}$

```

1:  $\text{success}_{\text{rate}} \leftarrow 0, \text{failure} \leftarrow 0, \text{length} \leftarrow 0$ 
2: for  $i = 1, \dots, n_\omega$  do
3:   // Run in parallel, or sequential for early termination.
4:    $s_s \leftarrow s.\text{SampleConfigSpace}()$  // Sample from the
5:    $s_g \leftarrow g.\text{SampleConfigSpace}()$  //  $\mathcal{C}_{\text{free}}$  space.
6:    $\text{success} \leftarrow 0, \text{steps} \leftarrow 0, s \leftarrow s_s$ 
7:    $\text{length}_{\text{trial}} \leftarrow 0$ 
8:   while
9:      $\text{steps} < \mathcal{K}_\omega \wedge \|p(s) - p(s_g)\| > d_G \wedge p(s) \in \mathcal{C}_{\text{free}}$ 
10:    do
11:       $s_p \leftarrow s, a \leftarrow \pi(s)$ 
12:       $s \leftarrow D.\text{predictState}(s, a)$ 
13:       $\text{steps} \leftarrow \text{steps} + 1$ 
14:       $\text{length}_{\text{trial}} \leftarrow \text{length}_{\text{trial}} + \|s - s_p\|$ 
15:    end while
16:    if  $\|p(s) - p(s_g)\| < d_G$  then
17:       $\text{success} \leftarrow \text{success} + 1$ 
18:       $\text{length}_{\text{trial}} \leftarrow \text{length}_{\text{trial}} + \|p(s) - p(g)\|$ 
19:    else
20:       $\text{failure} \leftarrow \text{failure} + 1$ 
21:    end if
22:    if  $(1 - p_s) < \text{failure}/n_\omega$  then
23:      return False, 0, 0 // Not enough success, we can
24:      terminate.
25:    end if
26:     $\text{length} \leftarrow \text{length} + \text{length}_{\text{trial}}$ 
27:  end for
28:   $\text{length} \leftarrow \frac{\text{length}}{\text{success}}, \text{success}_{\text{rate}} \leftarrow \frac{\text{success}}{n_\omega}$ 
29:  return  $\text{success}_{\text{rate}} > p_s, \text{success}_{\text{rate}}, \text{length}$ 

```

---

parameterized noise models for actions, observations, and robot localization, which improves model training and robustness. Stepping is fast compared to full-physics simulations because our simulator is kinematic. This speeds up RL training and roadmap building.

Algorithm 2 describes the roadmap building procedure, where an RL agent is trained once, and used on several environments. While building a roadmap for each environment, we first sample the obstacle map to the given density and store all candidate edges. Two nodes that are within the RL policy range,  $d_\pi$ , are considered candidates. Next, we partition all the candidate edges into subsets for distributed processing. The PRM builder considers each candidate edge and adds it, along with its nodes, to the roadmap if the AddEdge Monte Carlo rollouts returns success above threshold.

**Algorithm 2;** PRM-RL Build Roadmaps.

---

**Input:** Obstacle maps:  $[m_1, \dots, m_n]$   
**Input:**  $\pi$ : RL agent's policy.  
**Input:**  $D$ : Generative model of system dynamics.  
**Input:**  $\rho_\omega$ : Sampling density.  
**Input:**  $d_\pi$ : Policy range.  
**Input:**  $n_p$ : Number of processors.  
**Input:**  $p_s \in [0, 1]$ : Success threshold.  
**Input:**  $n_\omega$ : Number of attempts.  
**Output:** RL policy,  $\pi$ , Roadmaps,  $[\text{roadmap}_1, \dots, \text{roadmap}_n]$

```

1: Train RL agent with [12] given  $D$  as described in
2: Section IV-A.
3: for  $m$  in  $[m_1, \dots, m_n]$  /* In parallel for each env. */ do
4:   Sample environment map  $m$  with density  $\rho_\omega$  and store
5:   candidate edges w.r.t.  $d_\pi$ .
6:   Partition candidate edges in  $n_p$  subsets,  $[e_1, \dots, e_{n_p}]$ .
7:   for  $\text{edges}$  in  $[e_1, \dots, e_{n_p}]$  /* In parallel over workers. */ do
8:     for  $e$  in  $\text{edges}$  /* In parallel over threads. */ do
9:       if AddEdge: Run Alg 1 with  $\pi$ . then
10:        Add nodes if not in  $\text{roadmap}$ .
11:        Add edge  $e$  to the  $\text{roadmap}$ .
12:      end if
13:    end for
14:  end for
15: end for
16: return RL policy,  $\pi$ , Roadmaps,
17:  $[\text{roadmap}_1, \dots, \text{roadmap}_n]$ 

```

---

### C. Navigation

Finally, Algorithm 3 describes the navigation procedure, which takes a start and a goal position. These are added to the roadmap if not present. Then, the roadmap is queried for a list of waypoints. If no waypoints are returned, the algorithm returns the start and goal as the path to give the RL agent the opportunity to attempt to navigate on its own. In execution, a higher-level PRM agent gives the RL agent one waypoint at the time as a subgoal, clearing these goals sequentially as the RL agent gets within goal distance  $d_G$ , until the final destination is reached or  $\mathcal{K}_\omega$  is exceeded.

## V. RESULTS

We evaluate PRM-RL's performance on both floorplan maps and SLAM maps with respect to comparable baselines, construction parameters, simulated noise, and dynamic obstacles, as well as with experiments on robots. Section V-A describes the robot and training setup, evaluation environments, and baselines. Section V-B demonstrates PRM-RL's superior performance on floormaps with respect to baselines, while the following sections examine PRM-RL's characteristics in more depth. Section V-C demonstrates PRM-RL's robustness to noise, and Section V-D explores quality and cost tradeoffs with success threshold and sampling density. Since one of our goals is to assess PRM-RL for real-world

**Algorithm 3:** PRM-RL Navigate.

---

**Input:** PRM *roadmap*.  
**Input:**  $\pi$ : RL agent's policy.  
**Input:**  $s, g \in \mathcal{C}_{\text{free}}$ : Start and goal.  
**Input:**  $d_G$ : Sufficient distance to the goal.  
**Input:**  $\mathcal{K}_\omega$ : Maximum steps for trajectory.

- 1: Add start and goal  $s, g \in \mathcal{C}_{\text{free}}$ , to *roadmap* if needed.
- 2: Query *roadmap* and receive list of waypoints  $[w_1, \dots, w_N]$ ,  $w_1 = s$ ,  $w_N = g$ .
- 3: **for**  $w$  in  $[w_2, \dots, w_N]$ , **do**
- 4:   Set  $w$  as a goal for the RL agent  $\pi$ .
- 5:   Set current state,  $c$  as start state  $s$ .
- 6:    $steps \leftarrow 0$
- 7:   **while**  $c$  is not within  $d_G$  from  $w$  **do**
- 8:     Apply action  $\pi(c)$ , and observe the resulting state as new current state  $c$ .
- 9:      $steps \leftarrow steps + 1$
- 10:    **if**  $c$  is in collision **then**
- 11:     **returnError**: Collision.
- 12:    **end if**
- 13:    **if**  $steps > \mathcal{K}_\omega$  **then**
- 14:     **return** Error: Timeout.
- 15:    **end if**
- 16:   **end while**
- 17: **end for**
- 18: **return** Success.

---

navigation, Sections V-E and V-F show PRM-RL's applicability to SLAM and large-scale maps, and Section V-G analyzes sim2real experiments on real robots. Finally, to demonstrate the extensibility of PRM-RL to new situations, Section V-H shows PRM-RL works well on simulated robots with dynamic constraints, Section V-I explores PRM-RL's robustness in the face of dynamic obstacles, and Section V-J compares PRM-RL's performance on a suite of maps used by a visual policy baseline.

### A. Methodology

1) *Robot Setup*: We use two different robot kinematic models, *differential drive* [43] and *simple car model* [43], [58]. We control both models with linear and angular velocities commanded at 5 Hz, receive goal observations from off-the-shelf localization, and represent both as circles with 0.3 m radius. The obstacle observations are from 2-D lidar data, as shown in Fig. 5, with a 220° field of view (FOV) resampled to 64 rays. Following [55], we use an observation trace of the last  $\theta_n$  frames to provide a simple notion of time to the agent, with  $\theta_{nCM} = 3$  and  $\theta_{nDD} = 1$  for the car model and diff-drive, respectively. We use Fetch robots [53] for physical experiments.

2) *Obstacle-Avoidance Local Planner Training*: We train P2P agents with AutoRL over DDPG [12] with reward and network tuning for the differential drive robot, and reward tuning for the car robot. In both cases, the true objective for training the local planner is to navigate within 0.25 m of the goal, allowing the RL agent to cope with noisy sensors and dynamics. Table I depicts learned reward hyperparameters. DDPG actor and critic

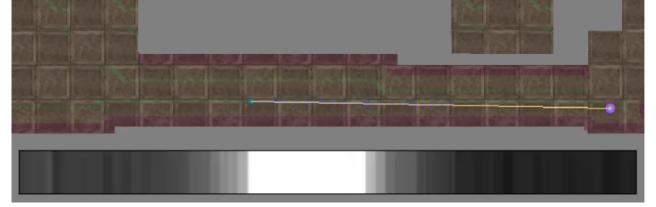


Fig. 5. Lidar observation that the robot uses for navigation. The background image shows a hallway with a clear path ahead of the robot and walls to the left and right. The inset shows the lidar image: white regions indicate no obstacles within 5 m, and progressively darker regions indicate closer and closer obstacles.

TABLE I  
P2P REWARD COMPONENTS AND THEIR AUTO RL-TUNED VALUES

Reward	Description	Min	Max	Diff Drive	Car Model
$r_{\text{goal}}$	1 when goal reached and 0 otherwise	0.0	100.0	<b>62.0</b>	<b>0.82</b>
$r_{\text{goalDist}}$	Negative Euclidean distance to goal	0.0	1.0	<b>0.38</b>	N/A
$r_{\text{goalProg}}$	Delta of Euclidean distance to goal	0.0	5.0	N/A	<b>2.03</b>
$r_{\text{collision}}$	1 on collision and 0 otherwise	-100.0	0.0	<b>-57.90</b>	<b>-1.80</b>
$r_{\text{clearance}}$	Distance to closest obstacle	0.0	1.0	<b>0.67</b>	N/A
$r_{\text{step}}$	Constant per-step penalty	-1.0	0.0	<b>-0.43</b>	<b>-0.10</b>
$r_{\text{turning}}$	Negative angular speed	0.0	1.0	<b>0.415</b>	N/A
$r_{\text{backward}}$	Negative backward speed	-1.0	0.0	N/A	<b>-0.64</b>

TABLE II  
ENVIRONMENTS

Environment	Type	Dimensions	Visual
Training	Floor map	23 m by 18 m	Fig. 3a
Building 1	Floor map	183 m by 66 m	Fig. 3d
Building 2	Floor map	60 m by 47 m	Fig. 3c
Building 3	Floor map	134 m by 93 m	Fig. 3b
Building Complex	Floor map	288 m by 163 m	Fig. 2
Physical Testbed 1	SLAM	50 m by 68 m	Fig. 6a
Physical Testbed 2	SLAM	203 m by 135 m	N/A (private)
Physical Testbed 3	SLAM	22 m by 33 m	Fig. 11c

are feed-forward fully connected networks. Actor networks are three layers deep, while the critics consists of a one or two-layer observation networks joined with the action networks by two fully connected layers. Actor, critic joint, and critic observation layer widths are  $(241, 12, 20) \times (607, 242) \times (84)$  for the reward-and-network trained differential drive model and  $(50, 20, 10) \times (10, 10) \times (50, 20)$  for the reward-trained car model. Appendix A contains the training hyperparameter details. The training environment is 14 m  $\times$  17 m, as shown in Fig. 3(a). To simulate imperfect real-world sensing, the simulator adds Gaussian noise,  $\mathcal{N}(0, 0.1)$ , to its observations.

3) *Evaluation Environments*: Table II lists our simulation environments, all derived from real-world buildings. *Training*, *Building 1–3*, depicted in Fig. 3, and *Building Complex* (see Fig. 2) are metric maps derived from real building floor plans. They are between 12 to 200 times larger than the training environment by area. *Physical Testbed 1*, as shown in Fig. 6, *Physical*

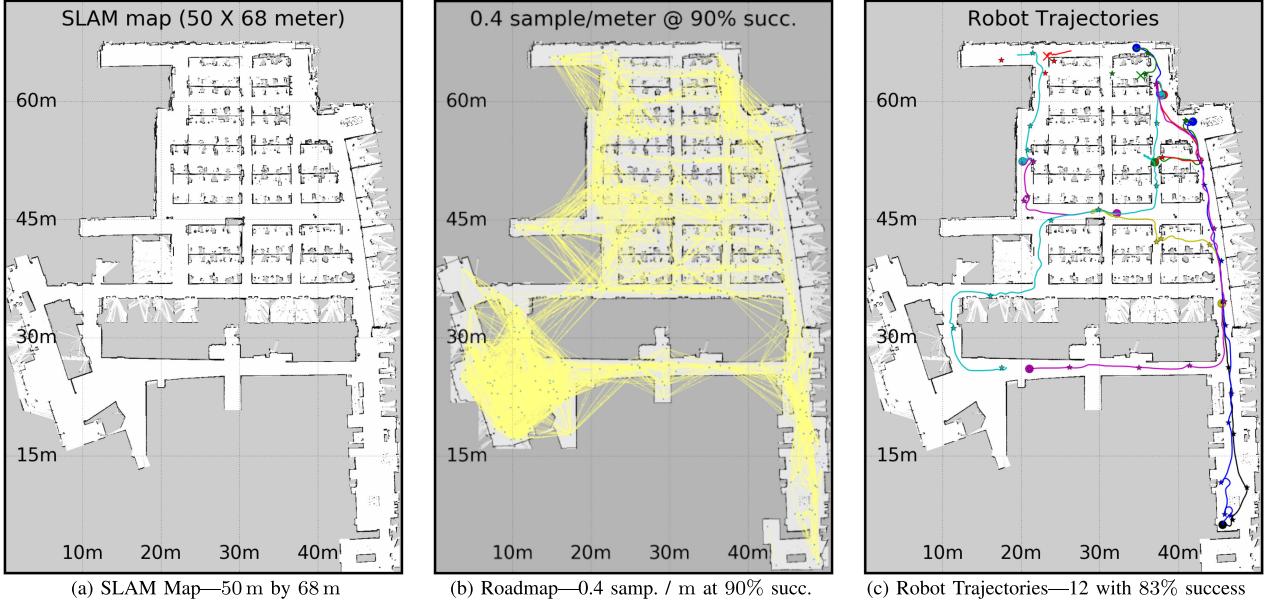


Fig. 6. PRMs for robot deployments are built over SLAM maps of the target environment. (a) SLAM map collected at the site of a robot deployment. (b) PRM-RL learns the effective connectivity of the map for the policy. (c) PRM-RL achieves 83% success on the robot.

TABLE III  
BASELINES

Label	Local Planner (LP)	Execution Policy	Obstacle Avoid LP	Avoid Execution	Monte Carlo rollouts	Description
AutoRL	N/A	AutoRL	N/A	Yes	No	Local AutoRL policy like [12] w/o guidance by a PRM.
PRM-SL	Straight Line	Straight Line	No	No	No	Straight-line PRMs [39] w/ straight-line execution policy.
PRM-GAPF	Straight Line	Guided APF	No	Yes	No	Straight-line PRMs [39] executed by guided APF like [11].
PRM-DWA	Straight Line	Guided DWA	No	Yes	No	Straight-line PRMs [39] executed by guided DWA [24].
<b>PRM-RL</b>	<b>AutoRL</b>	<b>AutoRL</b>	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>	<b>PRM-RL w/ AutoRL for roadmap &amp; execution (ours).</b>
PRM-HTRL	DDPG	DDPG	Yes	Yes	Yes	PRM-RL w/ DDPG for roadmap & execution ([23]).
SF	N/A	SF	N/A	Yes	No	Successor Features (SF) visual local planner [70].

*Testbed 2 and Physical Testbed 3*, as shown in Fig. 11(c), are derived from SLAM maps used for robot deployment environments.

4) *Roadmap Building*: For simplicity, we use uniform random sampling to build roadmaps. We connect PRMs with a  $p_s$  effective threshold of  $\geq 90\%$  over 20 attempts, with a max connection distance  $d_c$  of 10 m based on the effective navigation radius  $d_\pi$  for our P2P agents per [12], except where otherwise stated.

5) *Baselines*: Table III shows four selected baselines. The baselines differ in the local planner, used for building the roadmap, and the execution policy, which guides the robot. We select baselines given their ability to avoid obstacles and deal with stochasticity. Recall that PRM-RL relies on a stochastic policy capable of obstacle-avoidance to connect nodes in the roadmap using Monte Carlo rollouts of the policy.

The baselines for experimental comparison include a local planner based on AutoRL [12], PRM-SL [39], PRM-guided artificial potential field (GAPF) (a modification of [11]), and PRM-dynamic window avoidance (DWA). PRM-SL [39] uses roadmaps built with a straight-line planner and a straight-line execution policy. PRM-GAPF uses PRMs built with a straight-line planner, and an execution policy of APF, an artificial potential

field planner [40], guided by the PRM-SL path, similar to [11]. PRM-DWA is similar to PRM-GAPF using DWA, a dynamic window avoidance local planner [24].

In addition, we compare against two other baselines numerically: PRM-hand tuned reinforcement learning (HTRL) [23] is our original PRM-RL with hand-tuned DDPG as the planner; where not otherwise specified, PRM-RL refers to our current approach. Successor Features (SF) [70] is a visual-based navigation approach using discretized actions. We do not compare PRM-RL with RRTs here because this work focuses on solving the multiquery problem, while RRTs solve single-query problems, making them comparatively expensive for building long-range trajectories on the fly, especially when incorporating end-to-end controls. While large roadmaps can be expensive to construct, they can be reused for many queries across many robots, and queries are fast. For example, RRT solves a single query in the same environment in about 100 s [13], while PRM-RL produces a path in less than a second for a prebuilt roadmap (see Table IV), although the roadmap takes hours to build.

For comparisons to baselines, unless otherwise stated, each roadmap is evaluated on 250 queries selected from the  $\mathcal{C}_{\text{free}}$  between start and goal positions from 1.5 to 100 m. We measure start to goal distance by the shortest feasible path as estimated

TABLE IV  
PRM-RL PERFORMANCE VERSUS BASELINES

Baseline	PRM Density	Succ. %	99% Conf.	PRM-RL Nav.Δ%	Path Dist. (m) Mean	Path / Shortest. Std.	Clearance (m) Mean	Time (s) Plan. Exec.	Roadmap Nodes	Roadmap Edges	Coll Checks (x10 <sup>6</sup> )				
AutoRL	N/A	8.53	1.52	83.20	<b>14.2</b>	<b>15.9</b>	1.98	5.98	0.43	0.57	<b>0.03</b>	42.4	N/A	N/A	N/A
PRM-SL	Sparse	12.40	1.79	79.33	28.6	26.2	1.16	0.63	0.28	0.14	0.07	16.7	<b>1321</b>	28984	<b>1.4</b>
PRM-SL	Dense	9.47	1.59	82.27	20.3	17.3	<b>1.10</b>	<b>0.48</b>	0.28	0.13	0.43	<b>12.8</b>	3302	186491	9.0
PRM-GAPF	Sparse	29.20	2.47	62.53	57.2	42.5	1.35	0.70	0.30	0.13	0.06	44.6	<b>1321</b>	28984	<b>1.4</b>
PRM-GAPF	Dense	28.13	2.45	63.60	47.0	35.9	1.36	0.65	0.30	0.13	0.39	38.1	3302	186491	9.0
PRM-DWA	Sparse	40.53	2.67	51.20	277.6	192.6	7.23	17.09	<b>0.48</b>	0.11	0.06	167.1	<b>1321</b>	28984	<b>1.4</b>
PRM-DWA	Dense	40.13	2.67	51.60	266.2	215.7	6.44	8.20	0.47	<b>0.10</b>	0.40	178.8	3302	186491	9.0
PRM-RL	Sparse	86.53	1.86	5.20	61.5	36.7	1.16	1.90	0.43	0.18	0.14	70.4	<b>1321</b>	55270	74.0
PRM-RL	Dense	<b>91.73</b>	<b>1.50</b>	Best	60.0	36.5	<b>1.10</b>	1.82	0.42	0.14	0.75	62.7	3302	<b>332399</b>	360.0

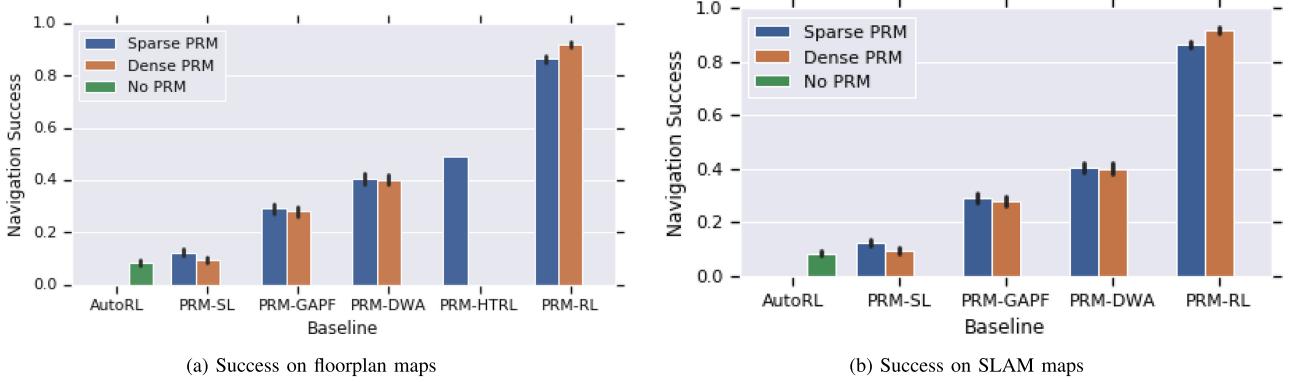


Fig. 7. PRM-RL outperforms the AutoRL local planner as well as PRM-SL, PRM-guided versions of GAPF and DWA, and our prior work PRM-HTRL. (a) PRM-RL's success is up to 93% on floorplan maps. (b) PRM-RL's success is up to 89% on SLAM maps.

by our simulation framework using a queen's-move discretized A\* search with adequate clearance for the robot to travel without collision.

### B. PRM-RL Performance on Floorplan Maps

Table IV shows PRM-RL's performance compared to baselines on Buildings 1-3 of our floorplan maps, using both sparse and dense PRMs; Fig. 7(a) also shows for reference our prior work PRM-HTRL. PRM-RL's average success rate in the dense condition is 91.7% over all three maps, which outperforms the baselines by 83.2% for pure AutoRL, 82.2% for dense PRM-SL, and 51.6% for dense PRM-DWA. Outperforming AutoRL's nonguided local policies is not surprising as they do not have the knowledge of the obstacle map, but we include it for completeness. Successful paths executed by AutoRL are shortest, but dense PRM-SL and dense PRM-RL are within 10% of the shortest feasible path as estimated by our simulation framework. DWA path lengths are much longer because DWA exhibits safe looping behavior in box canyons given the action space and noisy observations, but DWA maintains the best clearance to obstacles.

Analyzing collision checks (the complexity of building a roadmap), planning time, and execution time reveals interesting tradeoffs. First, PRM-RL requires 1–2 orders of magnitude more collision checks than PRM-SL; increasing map density only consistently helps PRM-RL. This is because of the Monte Carlo rollouts. At runtime, path finding requires no collision checking, as it performs a graph search on a prebuilt roadmap. At runtime, planning with a roadmap requires no collision checks, though local control policies may carry out collision checks in execution, e.g., DWA collision checks trajectories in velocity

space. Second, PRM-RL planning time is up to twice as long as PRM-SL, because the PRM-RL roadmaps contain more edges due to the RL local planner connecting to nodes that are not in the clear line of sight. Still, planning takes less than a second even for the densest maps, an insignificant part of execution time in both simulation and robot deployment. Finally, the execution time for PRM-RL is almost five times longer than PRM-SL because PRM-RL succeeds at longer trajectories thanks to RL control policy that adapts on-the-go to uncertainties not present in planning time, i.e., moving obstacles or sensor noise.

We can draw three observations from these results. First, guiding a local planner with PRM-RL can vastly improve the planner's success rate; this is not surprising as our local planners are not designed to travel long distances. Second, PRM-RL successfully enables local planners to transfer to new environments: the AutoRL policy only saw the Training Map in training, yet performs at a 91% success rate on our evaluation maps. Third, the PRM success rate is correlated with the success of the local planner. PRM-RL with an AutoRL policy and a sparse PRM build achieves 86.5% success, a 37.5% increase over the sparse 49% success rate reported in PRM-HTRL. This is evidence that investing in a high-fidelity local planner increases PRM-RL's performance of the overall navigation task. In contrast, moving to a denser PRM map (which we discuss in more detail in Section V-D) provides a lesser increase of 5.2%; this is still significant for deployment, however, as it also represents a 38.6% decrease in errors.

### C. PRM-RL Robustness to Noise

Sensors and actuators are not perfect, so navigation methods should be resilient to noise. Fig. 8 shows the evaluation of

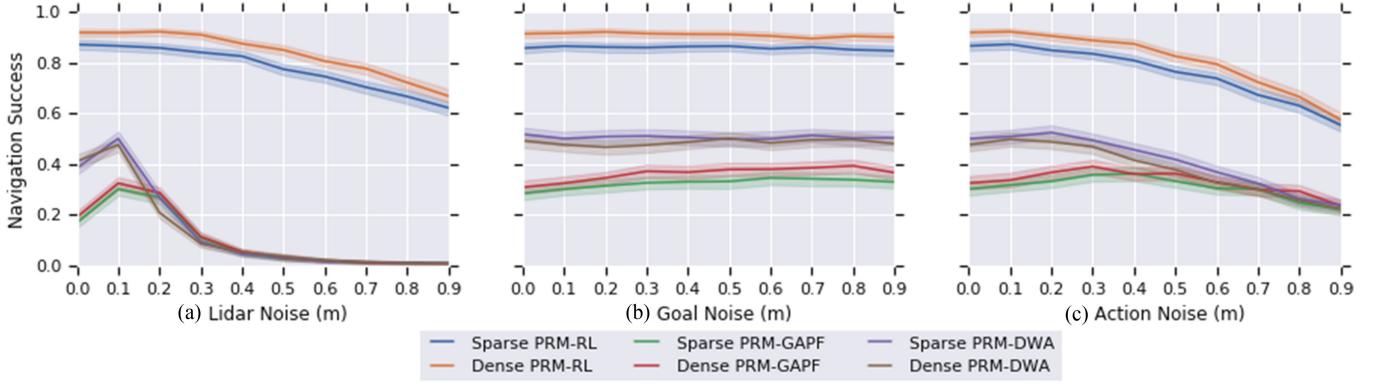


Fig. 8. PRM-RL is more robust to noise than PRM-GAPF or PRM-DWA. (a) As lidar noise increases, PRM-RL degrades slowly, showing a 28% drop at a noise level of 1 m, whereas PRM-GAPF and PRM-DWA degrade quickly to roughly 1% performance at noise of 0.75 m. (b) All methods show resistance to position noise (modeled as goal uncertainty). (c) As action noise increases, PRM-RL degrades slowly, showing a 37% drop at noise of 1 m; PRM-GAPF and PRM-DWA are degraded up to 39% and 54% of their peak performance, respectively.

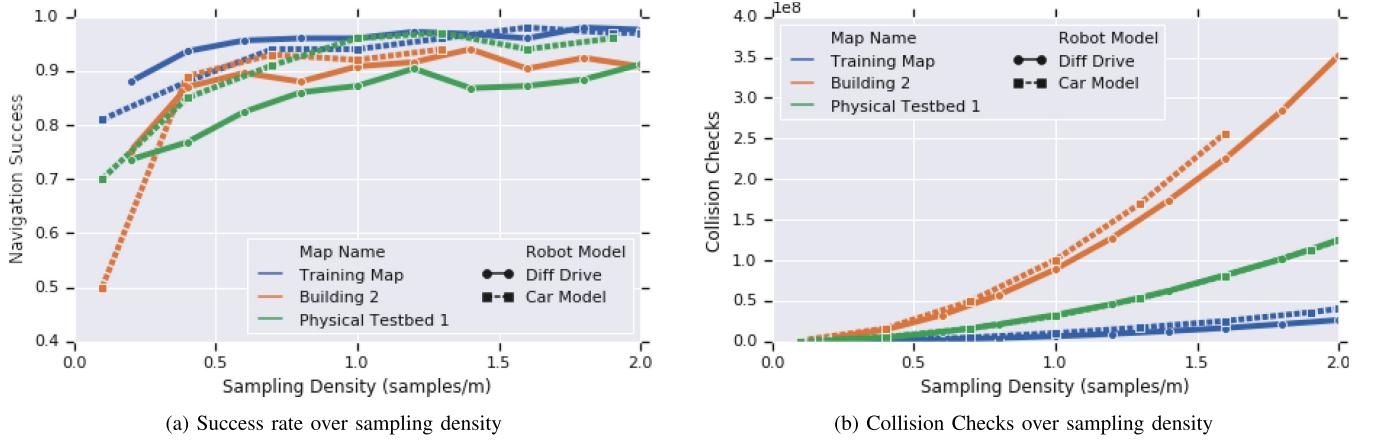


Fig. 9. Increasing sampling density improves performance at the cost of roadmap construction time. (a) As density increases, RL agents guided by PRM-RL succeed more often with a sweet spot of 1 node per square meter. (b) Cost rises prohibitively as sampling rises; over 1 node/m<sup>2</sup>, collision checks for the training map surpass our largest floorplan roadmap collected at 0.4/m<sup>2</sup>.

PRM-RL, PRM-GAPF, and PRM-DWA on the Training map, Buildings 1–3 and Physical Testbed 1 over simulated Gaussian noise sources with mean 0.0 and  $\sigma$  in the following ranges.

- 1) *Lidar sensor noise*  $\sigma_l$  from 0 to 0.9 m, over three times the radius of the robot.
- 2) *Goal position noise*  $\sigma_g$  from 0 to 0.9 m, over three times the radius of the goal target.
- 3) *Action noise* of velocity  $\sigma_v$  0 to 0.9 m/s and angular velocity  $\sigma_a$  0.9 rad/s.

As lidar and action position noise increase, PRM-RL shows only a slight degradation of 28% on lidar noise and 37% on action noise, even at 0.9 m. In contrast, PRM-GAPF and PRM-DWA degrade steeply with respect to lidar noise, with success rates dropping to less than 1%. These methods seem to be more resistant to increased action noise, but still drop to 39% and 54% of their peak scores for PRM-GAPF and PRM-DWA, respectively. All methods were relatively resistant to goal noise, with less than 10% falloff. PRM-RL outperformed PRM-GAPF and PRM-DWA in all conditions.

PRM-RL is resilient to lidar, localization, and actuator noise on the order of tens of centimeters, which is larger than the typical lidar, localization, and actuator errors we observe on our robot platforms, indicating that PRM-RL is a feasible approach to deal with the kinds of errors our robots actually encounter. This is similar to the trend to noise sensitivity reported in [12], and suggests that overall navigation resilience to noise is correlated to that of the local planner.

#### D. Impact of Sampling Density and Success Thresholds

To deploy PRM-RL on real robots, we need to choose sampling density and success threshold parameters that provide the best performance. Fig. 9 shows that PRM-RL success rate increases steadily up to a sampling density of 1.0 samples/m, which is roughly twice the size of our robot, and, then, levels off. At the same time, collision checks increase rapidly with sampling density; we have observed that over 1.0 samples/m, the roadmap for the training environment requires more collision

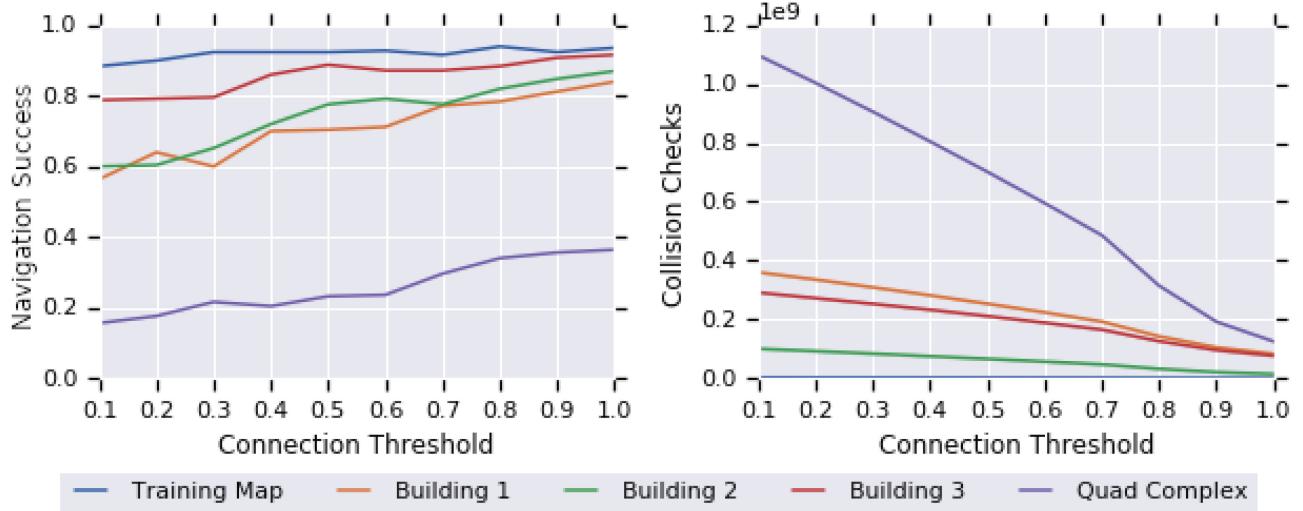


Fig. 10. Increasing required edge connection success improves performance and reduces collision checks. (a) As the threshold for connecting nodes in the PRM rises, RL agents guided by PRM-RL succeed more often with a sweet spot of 90% and higher. (b) Furthermore, early termination enables PRM-RL to skip unneeded connectivity checks for savings exceeding 60%, an effect important on large roadmaps.

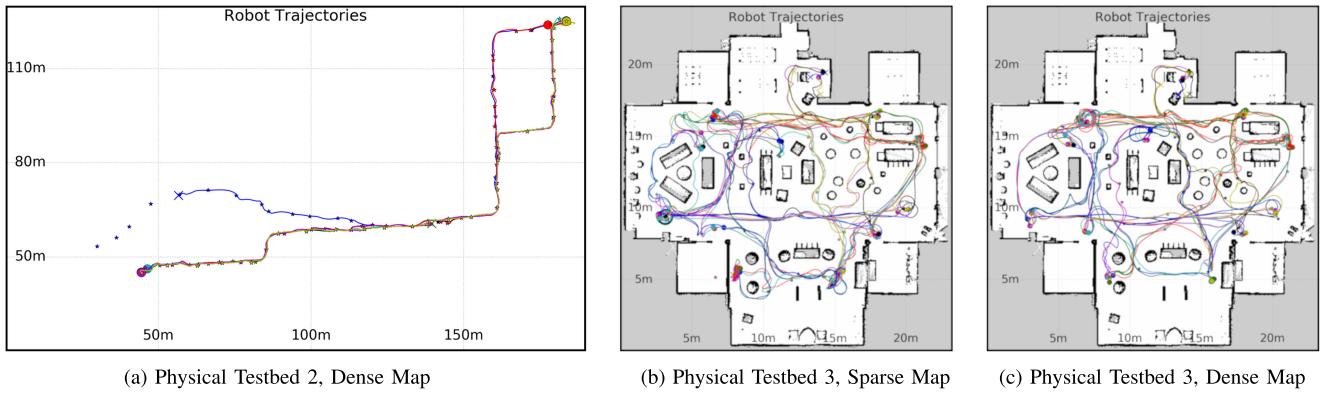


Fig. 11. Trajectories collected from PRM-RL execution on real robots. Trajectories are in color, circles represent successful navigation, X's represent emergency stops. (a) Several queries executed on a differential drive robot and tracked with onboard localization in a real office environment in *Physical Testbed 2*. The longest successful single trajectory was 221.3 m. The floorplan and PRM connectivity are not displayed for privacy. (b)–(c) *Physical Testbed 3* is a simulated living room, where 128 trajectories collected with sparse and dense map densities. The denser map achieved 6% higher success rate and produced on average 2.5 m shorter paths.

checks than some of the larger evaluation roadmaps collected at 0.4 samples/m. While PRM-SL theory predicts that performance would continue to improve with increased sampling [38], [33], this suggests that beyond a critical density PRM-RL performance is robot-noise-bound, and that sampling beyond this density provides little benefit at a rapidly increasing cost.

These experiments evaluate global planning; however, note the average distance of points at 2.0 samples/m<sup>2</sup> is roughly 0.5 m, the true objective distance used to train our AutoRL local planners. This suggests the asymptotic behavior in these experiments could be explained by the PRM sampling density approaching a distance where the local planner can almost always find a path to a nearby PRM node. Also note the car model's success rate differs from the diff-drive model, suggesting that success rate is model bound rather than map bound.

Fig. 10 shows that PRM-RL's success rate increases with the required connection success rate. Because our connection algorithm terminates earlier for higher thresholds when it detects failures, collision checks drop as the success threshold rises. At the end, for larger roadmaps, the success threshold of 100% not only produces the most reliable roadmaps, but requires fewer collision checks to build them.

These results suggest choosing map densities up to 1.0 samples/m with as high a success connection threshold as possible. In this article, we compare sparse and dense parameters: a sampling density of 0.4 samples/m and an effective success connection threshold of  $\geq 90\%$ , which enables comparison with [23], and a density 1.0 samples/m with a threshold of 100%; in almost all evaluations, we observe better performance with the dense parameters.

### E. PRM-RL Performance on SLAM Maps

Floorplans are not always available or up-to-date, but many robots can readily generate SLAM maps and update them as buildings change. To assess the feasibility of using SLAM maps, we evaluate PRM-RL on SLAM maps generated from some of the same buildings as our floorplan roadmaps. Fig. 6 illustrates a sample roadmap built from a SLAM map generated with the ROS distribution of the GMapping algorithm [27] with the default resolution of 5 cm per pixel. This map corresponds to *Physical Testbed 1*, part of floor 2 of Building 1 in Fig. 3(d) and the upper center section of the large-scale roadmap in Fig. 2. This SLAM-derived roadmap has 195 nodes and 1488 edges with 2.1 million collision checks.

We compare PRM-RL with our baselines on the three Physical Testbed SLAM maps. PRM-RL’s success rate is 89% on the dense PRM, a 97% relative increase over PRM-DWA and a 157% increase over PRM-GAPF.

These results show that the performance of PRM-RL with an AutoRL policy is weak but comparable to its performance on floorplan maps, and exceeds all other baselines; it is even superior to PRM-HTRL on floorplan maps. These results indicate PRM-RL performs well enough to merit tests on roadmaps intended for real robots at physical sites, which we discuss in the following two sections.

### F. Scaling PRM-RL to Large-Scale Maps

Our robot deployment sites are substantially larger than our simulated test maps, raising the question of how PRM-RL would scale up. For example, the SLAM map discussed in the previous section is only part of one building within a quad-building complex. Where the SLAM map is  $78 \times 44$  m, a map of the quad-building complex is  $288 \times 163$  m. To assess PRM-RL’s performance on large-scale maps, we build and test roadmaps for maps covering all deployment sites.

Fig. 2 depicts a large floorplan roadmap from the quad-building complex. This roadmap has 15 900 samples and 1.4 million candidate edges prior to connection attempts, of which 689 000 edges were confirmed at a 90% success threshold. This roadmap took 4 days to build using 300 workers in a cluster, and required 1.1 billion collision checks. PRM-RL successfully navigates this roadmap 57.3% of the time, evaluated over 1000 random navigation attempts with a maximum path distance of 1000 m. Note our other experiments use a maximum path distance of 100 m, which generally will not cross the skybridge in this larger map. For reference, using our default evaluation settings, PRM-RL navigates this roadmap 82.3% of the time.

For our other candidate robot deployment site, we use a large SLAM map,  $203 \times 135$  m. We constructed a roadmap with 2069 nodes and 53 800 edges, collected with 42 million collision checks at the higher success threshold of 100%. PRM-RL successfully navigated this 58.8% of the time, evaluated over 1000 random navigation attempts. As on our smaller SLAM map, the failure cases indicate that the more complex geometry recorded by SLAM proves problematic for our current policies.

These results indicate that PRM-RL’s simulated performance on large-scale roadmaps surpasses the average success threshold

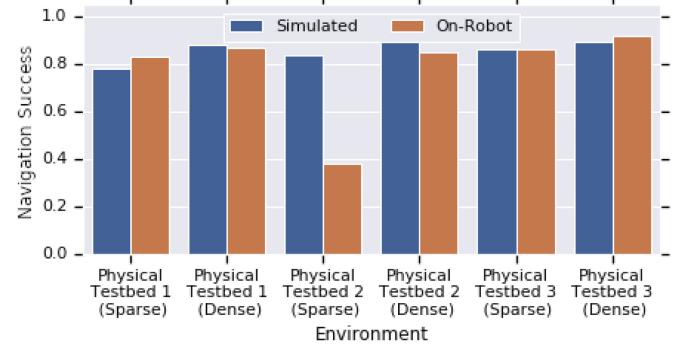


Fig. 12. PRM-RL closes the sim to real gap. Performance in simulation and on robots is similar, despite an e-stop policy that stops robots more aggressively than in simulation. Note that on Physical Testbed 2, a dense PRM overcomes an obstacle that thwarted a sparse PRM.

we observed previously in [23], making it worthwhile to test on real robots at the physical sites.

### G. Transfer of PRM-RL to Physical Robots

We empirically evaluate PRM-RL on three physical environments on two differential-drive robots. First, we evaluate PRM-RL with the AutoRL policy for a differential-drive robot in *Physical Testbed 1*, as shown in Fig. 6(a). We collected 27 trajectories over 831 m of travel with an overall success rate of 85.2%; the longest successful trajectory was 83.1 m. Fig. 6(c) shows the trajectories of 12 runs.

Second, we evaluate PRM-RL in *Physical Testbed 2*. For a variant of the roadmap generated at 90% success rate with a density of 0.4 samples/m, we collected eight trajectories over 487.2 m of travel; the longest successful trajectory was 96.9 m. We cannot directly compare this evaluation to our simulated runs because the e-stop policies designed to protect the robot do not match our simulation termination conditions. Nevertheless, we recorded three successful runs out of eight, a 37.5% success rate. We, then, tested a variant of the roadmap generated at 100% success rate and a density of 1.0 samples/m over 13 runs on 2542.1 m of travel for an improved 84.6% success rate, shown in Fig. 11(a); the longest successful trajectory was 221.3 m.

Third, we apply the same evaluation parameters from *Physical Testbed 2* onto a novel environment, *Physical Testbed 3*. We collected 128 runs, 64 for each variant of the roadmap, testing both with the same sets of start/goal points. The denser map achieved a 92.2% success rate compared to 85.9% for the sparse map. Since *Physical Testbed 3* contains large sections of free space, using a denser map resulted in more optimized node connections and reduced robot traversal by an average of 2.5 m per run.

Fig. 12 summarizes these results. Despite more aggressive episode termination policies on robots (near-collisions are treated as failures), we nonetheless observe similar results: over several different roadmaps constructed at different densities and success criteria, PRM-RL achieves 85.8% success on robots with an average sim2real gap of 7.43%. These results show that the performance of PRM-RL on robots is correlated with the performance of PRM-RL in simulation. This makes PRM-RL a

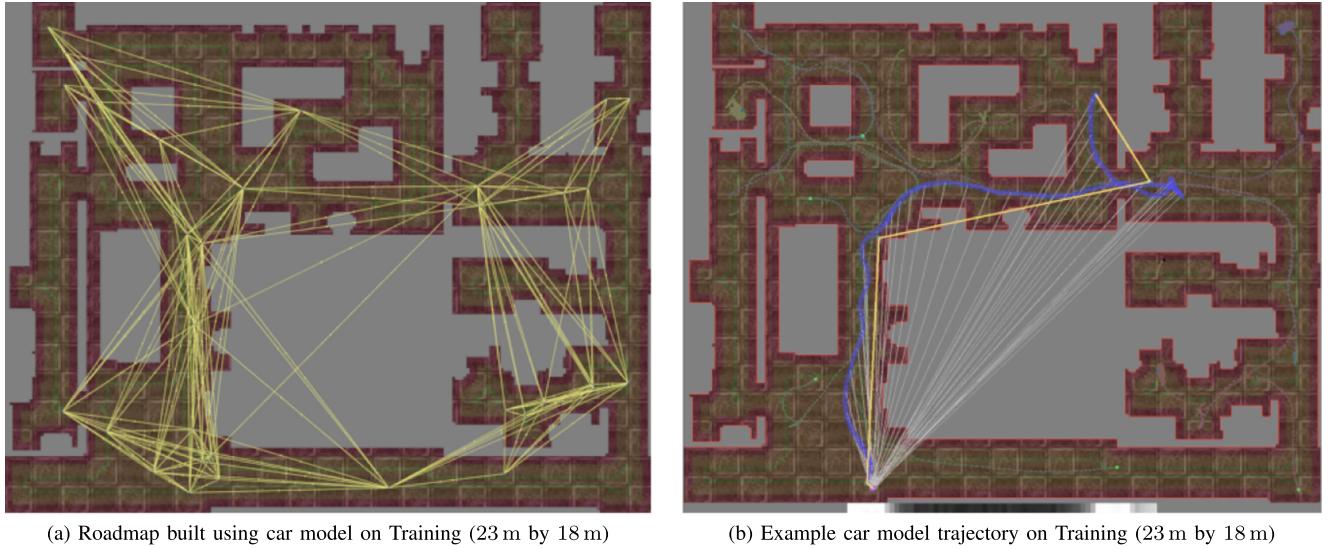


Fig. 13. PRMs can be built for agents with dynamic constraints. (a) Roadmap built for our training environment using a nonholonomic car model. (b) Example car model trajectory; the upper right shows a 3-point turn to change the robot orientation. Yellow lines are the PRM path, the blue line is the agent's trajectory, white lines indicate progress toward the goal, and light green dots represent previous evals.

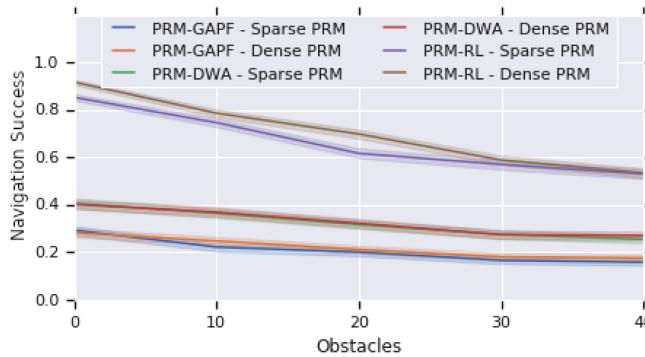


Fig. 14. PRM-RL is resilient in the presence of dynamic obstacles. With 40 moving agents, performance degrades to 53.1%, a higher score than PRM-GAPF and PRM-DWA achieve in the default condition.

useful method for closing the sim-to-real gap, as improvements or regressions in simulation are likely to be reflected in performance on robots.

#### H. PRM-RL With Kinodynamic Constraints

To demonstrate that PRM-RL could guide more complicated robots, we develop a drive model for a simulated F1/10 car [18] with dynamics following [58] with velocity and acceleration constraints enforced by our simulator. In this setup, AutoRL learns a steering function for a robot with kinodynamical constraints, effectively learning to respect velocity and acceleration constraints. Algorithm 1 encodes those constraints globally by connecting only reachable nodes, making PRM-RL effectively a kinodynamic planner [14]. Average success over the four maps in Fig. 3 is 85.8% with a standard deviation of 1.0%; average success in simulation on *Physical Testbed 1* with a goal distance of 0.25 m is 85.8%. Fig. 13(a) illustrates a roadmap built with this model over our training map with 0.4 samples/m connected

with a 90% success rate; this roadmap has 32 nodes and 313 edges connected after 403 000 edge checks. On this roadmap, PRM-RL exhibits an 83.4% success rate, including cases where the car needs to make a complex maneuver to turn around, as shown in Fig. 13(b). These results are comparable to results on the robot, indicating that PRM-RL is a viable candidate for further testing on more complicated robot models.

#### I. PRM-RL With Dynamic Obstacles

PRM-RL is also resilient in the face of dynamic obstacles, relying on the local planner to avoid them without explicit replanning. We simulated pedestrians with the social force model [31] and tested PRM-RL, PRM-GAPF, and PRM-DWA on Buildings 1–3 with 0–40 added agents (Fig. 14). While all methods showed a similar degradation in the presence of obstacles, on average 39.0%, PRM-RL's performance only dropped to 53.1% with 40 obstacles, superior to PRM-GAPF (28.7%) and PRM-DWA (40.3%) even in the zero-obstacle condition. Thus, a resilient local planner can enable PRM-RL to handle dynamic obstacles even though the framework has no explicit support for dynamic replanning.

#### J. PRM-RL in Synthetic Environments

The SFs visual navigation approach achieves 100% success on training maps and 98% success on transfer to new maps [70]. SFs navigates to single targets using vision and a discretized action space. In contrast, PRM-RL navigates to arbitrary targets using lidar and a continuous action space. Nevertheless, both approaches navigate in spaces qualitatively similar to each other in simulation and in testing on robots.

Therefore, we evaluate PRM-RL and our baselines in simulation on close approximations of the maps used in [70], which include four maze-like simulated maps and two physical testbed environments. Fig. 15 shows the results: Both AutoRL and

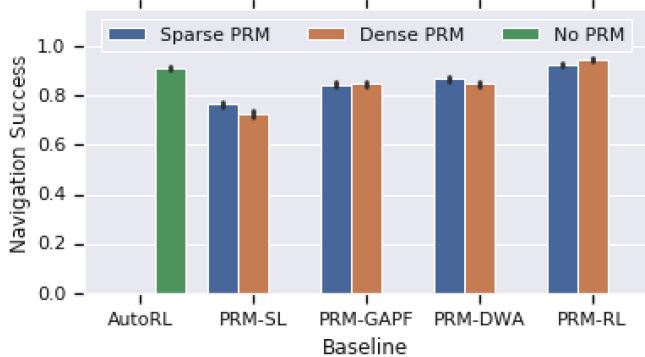


Fig. 15. PRM-RL is competitive with SFs [70]. On transfer maps, PRM-RL achieves over 94% success with noisy lidar, continuous actions, and arbitrary goals, whereas SFs reports 98% success with vision, discrete action, and a single goal.

PRM-RL achieved over 90% success on the synthetic maps, with PRM-RL on a dense map achieving 94.6% success. While SFs's 98% success is 3.4% higher than PRM-RL, it has access to a visual sensor, executes discretized actions, and navigates to a single target.

We can draw two conclusions from these findings. First, PRM-RL generalizes well to environments, which are different from both our training environment and our previous testing environments. Second, PRM-RL is competitive with methods specifically designed for other environments.

## VI. ANALYSIS

In the previous section, we empirically established correlations between the local planner's competence and PRM-RL's resilience to noise. We also explored the contributions of sampling densities, success thresholds, and obstacle map sources to the success of overall navigation. We concluded that 1) using an execution policy that is resilient to noise and avoids obstacles as a local planner improves the overall success of the hierarchical planner; 2) a success threshold of 100% improves overall navigation success, 3) the upper bound to navigation success is not dependent on density but policy performance and robot type, and 4) using realistic obstacle maps, such as SLAM maps, as a basis for building roadmaps provides simulation results closer to reality.

This section provides a deeper analysis of those empirical findings. Section VI-A analyzes the impact of local planner's obstacle avoidance and noise resilience on roadmap construction. Section VI-B examines the computational complexity of PRM-RL, Section VI-C discusses causes of failure for trajectories over multiple waypoints, and Section VI-D discusses limitations of the method and future work.

### A. PRM-RL Roadmap Connectivity

Unlike straight-line planners, RL agents can often go around corners and smaller obstacles; Fig. 16 shows how this effectively transforms the configuration space to make obstacles smaller. While the agent never traverses these corner points, as they are not in  $\mathcal{C}_{\text{free}}$ , they nevertheless do not serve to block the

agent's path, unlike central portions of a larger body, which might block or trap the control agent in a local minimum. If we model this as an effective reduction in radius of a circular obstacle  $f_\pi$  with respect to a policy  $\pi$ , and model the connection region as a disc filled with randomly sampled obstacles from 0% to 100% in total area density  $\rho_o$ , we can estimate an upper bound on connection success in the idealized case in which obstacles do not occlude and the chance of connection is just the complementary probability of encountering an obstacle over a region of space,  $1 - \rho_o$ . This corresponds to the looser bound  $1 - \rho_o f_\pi^2$  in the RL case. Therefore, a conservative estimate of the ratio of samples connected by PRM-RL to those connected by PRM-SL is

$$\frac{\text{conn}_{\text{PRM-RL}}}{\text{conn}_{\text{PRM-SL}}} = \frac{1 - \rho_o f_\pi^2}{1 - \rho_o}. \quad (4)$$

This simplified model indicates that it becomes harder to connect points as obstacle density increases, but PRM-RL has an increasing advantage over PRM-SL in connecting these difficult roadmaps as RL's ability to cut corners increases. Hence, in environments with dense obstacles, it makes sense to invest in execution policies that avoid obstacles really well, and use them as local planners. Alternately, it suggests that policies that can learn strategies for dealing with frequently occurring obstacle topologies, such as box canyons and corners, are a fruitful area for future work. Because the branching factor of roadmap planning is proportional to the points in the connectivity neighborhood, planning cost can increase exponentially with connection radius, and following [25], we set neighborhood size empirically to balance the benefits of finding points within the effective navigation radius of the planner against the drawbacks of a connectivity neighborhood, which contains so many points that planning becomes infeasible.

Conversely, PRM-RL does not connect nodes in a roadmap where a policy cannot navigate reliably. This is the key difference from PRM-SL, and is the cause for the upper limit on performance improvements as the roadmaps increase in density—the roadmaps are policy-bound, rather than sampling bound. One question to ask is why local control policies cannot learn to drive safe by repeatedly replanning the path (e.g., via A\* searches or variants). However, an analysis of how noise impacts the behavior of policies indicates that policies, which do not memorize the environment may overestimate their ability to navigate because the hazards that they can observe locally may not represent an accurate picture of the hazards of the global environment.

To see why, suppose a policy has learned to navigate safely in the presence of noise by maintaining a distance  $d_{\text{safety}}$  from walls. Modeling time as discrete and assuming the robot is traveling at a constant speed, so that on each time slice, the robot moves a constant distance  $d_{\text{step}}$  units forward, let us further model sources of variability as transient Gaussian noise orthogonal to the robot's travel  $\mathcal{N}_{\text{pos}}(0, \sigma_{\text{pos}})$  with zero mean and standard deviation  $\sigma_{\text{pos}}$ . This results in a probability of collision per step of  $\frac{1}{2} \text{erfc}(\frac{d_{\text{safety}}}{\sqrt{2}\sigma_{\text{pos}}})$  (the cumulative distribution function of the Gaussian noise model  $\mathcal{N}_{\text{pos}}$  evaluated at  $-d_{\text{safety}}$ , expressed in terms of the complementary Gauss error function  $\text{erfc}(x)$ ). Fig. 17(a) shows that when the robot is traveling in a narrow

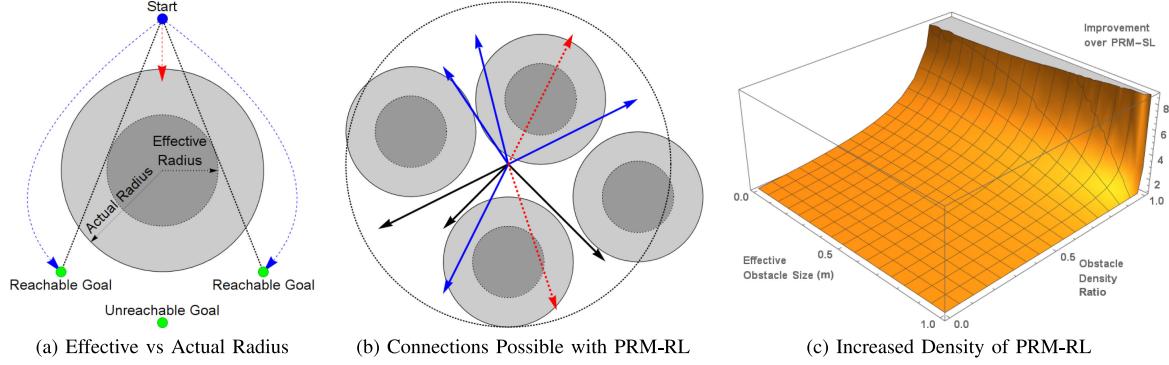


Fig. 16. PRM-RL enables capturing valid samples not visible by line of sight. (a) Ability of RL to go around corners makes obstacles effectively smaller in configuration space. (b) This means more connections can be made for a given connectivity neighborhood. Solid black arrows represent valid connections for either PRM-SL or PRM-RL, dotted red arrows represent invalid connections for either method, and blue arrows indicate valid trajectories recovered by PRM-RL. (c) Compared to PRM-SL, PRM-RL recovers many more potential connections as obstacles grow denser and RL gets better.

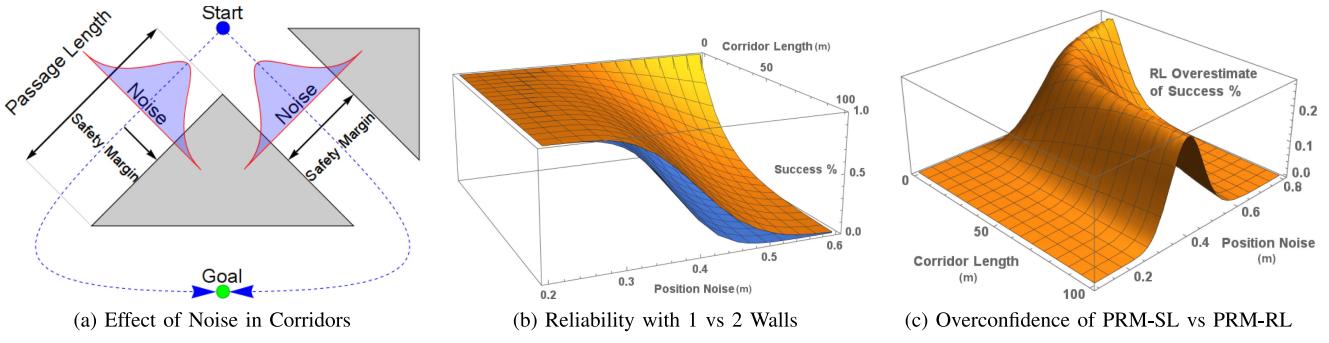


Fig. 17. PRM-RL enables capturing hazards in the environment difficult to learn with RL. (a) RL agents may learn to avoid obstacles, but not every location in the environment has identical clearance; on the left, the robot can hug one wall, but on the right, it must pass between two walls, so uncertainty in controls leads to two possible failure modes. (b) Therefore, as paths lengthen, the chance of navigating reliably drops faster in corridors than in wall hugging (blue region). (c) This leads to RL overconfidence in a regime between where both PRM-RL and RL are reliable and where they both are not; PRM-RL can encode this in the roadmap by deleting the right edge in (a).

corridor, it is twice as likely to collide as it does when hugging a wall, even though it may be maintaining  $d_{\text{safety}}$  from any given wall at all times. Over a path of length  $d_{\text{corr}}$ , a conservative lower bound on the chance of collision rises exponentially with the number of steps it takes to traverse the path

$$P_{\text{survival}} = \left(1 - \frac{1}{2} n_{\text{walls}} \text{erfc} \left( \frac{d_{\text{safety}}}{\sqrt{2} \sigma_{\text{pos}}} \right) \right)^{\frac{d_{\text{corr}}}{d_{\text{step}}}} \quad (5)$$

causing the narrow corridor case to become unsafe faster than the wall-hugging case as shown in the blue region of Fig. 17(b). This means that an RL policy that judges its safety based on locally observable features can overestimate the safety of a path in the region between where both PRM-RL and RL would succeed and where both PRM-RL and RL would fail [see Fig. 17(c)]. The same would be true of RL guided by PRM-SL based on clearance, such as in grid-based sampling [19] and Safety PRM [50]. In this case, RL or RL guided by PRM-SL can make erroneous navigation choices, whereas PRM-RL simply does not add that link to the roadmap. While in theory, an agent could be designed to cope with this specific issue, other scenarios can present similar problems: no local prediction of collision probability can give a true guarantee of trajectory safety. While this is an idealized analysis, our experience is that real-world

environments can be more pathological—for example, at one site, curtains behind a glass wall induced a persistent 0.5 m error in the robot's perceived location, causing it to drive dangerously close to the wall for long distances despite the presence of a safety layer. PRM-RL provides an automated method to avoid adding those unreliable connections in the roadmap, resulting in the roadmaps that more sparsely connect but transfer more reliably to the robot.

### B. PRM-RL Computational Complexity

In this section, we assess the computational complexity of building a PRM-RL roadmap with known RL policy range  $d_{\pi}$ , sampling density  $\rho_{\omega}$ , workspace volume  $V_{\mathcal{W}}$ , and connection attempts  $n_{\omega}$ . The cost of building a PRM with  $n$  nodes is dominated by the cost of two distinct steps. The first step identifies potential edges by finding  $m$  potential nearest neighbors for each of the  $n$  nodes. While approaches such as PRM\* can construct PRMs with  $O(n \log n)$  edge tests, achieving this performance requires limits on  $m$ , such as connecting only the  $m$  nearest neighbors or shrinking the radius of connection as  $n$  increases [36]. Approaches using a fixed connection radius, such as Simplified PRM [38], require  $O(n^2)$  edge tests [36]. The second step evaluates the validity of each of the  $n m$  candidate

edges by performing collision checks along each edge, and adds valid edges to the graph.

PRM-RL samples nodes in the workspace  $\mathcal{W}$  and attempts to connect all neighbors within range of the policy  $d_\pi$ , which can be determined empirically [12] and is independent of the workspace volume  $V_\mathcal{W}$ . Similarly, in Section V-D, PRM-RL's performance shows diminishing returns beyond a sampling density  $\rho_\omega$ , also independent of workspace volume. Capping sampling at density  $\rho_\omega$  makes the number of nodes a function of the workspace volume

$$n_w = V_\mathcal{W} \rho_\omega = O(d^{D_w} \rho_\omega) \quad (6)$$

where  $D_w \leq 3$  is the workspace dimensionality, and  $d$  is the radius of the smallest  $D_w$ -dimensional sphere that contains the workspace  $\mathcal{W}$ . However, because the neighbor volume is a function of policy range and workspace dimensionality,  $V_\pi \propto d_\pi^{D_w}$ , the number of neighbors

$$m_\pi = V_\pi \rho_\omega = O(d_\pi^{D_w} \rho_\omega) \quad (7)$$

is a variable independent from  $n_w$ , although their ratio is fixed for any given workspace. Let  $d_\pi = cd$  for some constant  $c \in \mathbb{R}$ . More capable policies have  $c$  closer to 1, while less capable policies have  $c$  closer to 0. For simplicity, we assume that  $0 < c \leq 1$ , when limiting the search radius, even if policy's radius might exceed the workspace's radius. Thus

$$\frac{m_\pi}{n_w} \propto \frac{d_\pi^{D_w} \rho_\omega}{d^{D_w} \rho_\omega} = c^{D_w} \leq 1. \quad (8)$$

Therefore, we analyze PRM-RL with respect to nodes  $n_w$  and neighbors  $m_\pi$  in terms of the source variables that determine them: workspace volume  $V_\mathcal{W}$ , sampling density  $\rho_\omega$ , and effective range  $d_\pi$ . First, the edge identification step is  $O(n_w \log n_w)$  because the number of nearest neighbors given in (7) is independent of the total number of nodes  $n_w$  and can be found in  $O(\log n_w)$  with efficient approximate nearest neighbor searches, as described in [36]. From there

$$O(n_w \log n_w) = O(d^{D_w} \rho_\omega \log d^{D_w} \rho_\omega), \text{ due to (6)} \quad (9)$$

$$= O(d^{D_w} \rho_\omega \log d \rho_\omega), D_w \text{ is constant.} \quad (10)$$

Second, for the edge validation phase, PRM-RL validates  $m_\pi * n_w$  candidate edges with  $n_\omega$  rollouts. To validate one edge rollout, PRM-RL performs on the order of  $O(d_\pi)$  collision checks. Therefore, the cost of adding neighbors to the roadmap is

$$O(n_w m_\pi n_\omega d_\pi) = O(d^{D_w} \rho_\omega^2 d_\pi^{D_w+1} n_\omega) \quad (11)$$

by substituting (6) and (7) and rearranging. Combining (10) and (11), we arrive at the total cost

$$O(d^{D_w} \rho_\omega \log d \rho_\omega + d^{D_w} \rho_\omega^2 d_\pi^{D_w+1} n_\omega) \quad (12)$$

$$= O(d^{D_w} d_\pi^{D_w+1} \rho_\omega^2 n_\omega) \quad (13)$$

because for a typical map  $\rho_\omega d_\pi^{D_w}$  dominates  $\log d \rho_\omega$ .

Equation (13) exposes the following power sources.

- 1) Complexity is  $O(d_\pi^{2D_w+1})$  in the policy range, so local planners that can reliably cover longer distances increase

the computational cost of the roadmap. We recommended choosing a shorter connection distance  $d_\omega < d_\pi$  even if the policy is capable of longer connections.

- 2) When the workspace is much larger than the reach of the policy ( $0 < c \ll 1$ ), the complexity is almost linear in workspace volume.
- 3) Complexity is linear in connection attempts  $n_\omega$ .
- 4) Complexity is quadratic in sampling density  $\rho_\omega$ , making it worthwhile to assess the limiting sampling density before building large numbers of roadmaps.

Each edge connection attempt is independent, so roadmap building can be parallelized up to the expected number of samples  $V_\mathcal{W} \rho_\omega$ . If parallel rollouts are performed instead of early termination, this can be parallelized further by an additional factor of  $n_\omega$ . Thus, given  $n_p \leq V_\mathcal{W} \rho_\omega n_\omega$  processors, the effective time complexity can be reduced up to  $O(\frac{V_\mathcal{W} \rho_\omega n_\omega}{n_p} \rho_\omega d_\pi^{D_w+1})$ , possibly alleviating some of the time cost of increased sampling.

Finally, note that when using early termination, increasing the success threshold  $p_s$  often (but not always) reduces the required number of connection attempts  $n_s$ . In the worst-case scenario where we require  $p_s = 0.5$ , early termination can at best cut  $n_s$  to  $\frac{n_\omega}{2}$ , but as  $p_s$  increases the number of failures needed to exclude an edge,  $\frac{n_\omega}{1-p_s}$ , drops toward 1. Conversely, if navigation is successful, then, the full  $n_\omega$  samples need to be collected for an edge; the distribution of successes and failures, thus, has a large effect on the cost. One way to control this cost is to reduce the max connection distance  $d_\omega$  to less than the effective policy navigation distance  $d_\pi$ ; in this case, the agent is more often expected to succeed, and  $n_\omega$  can potentially be reduced. We have observed that these tradeoffs can significantly affect the cost of a run, but must be studied empirically for the environments of interest.

### C. PRM-RL Trajectory Execution

Because PRM-RL construction calculates the probability of success before adding an edge, we can estimate the expected probability of success of a long-range path that passes through several waypoints. Recall that to add an edge to the roadmap, we collect  $n_\omega = 20$  Monte Carlo rollouts and require an observed proportion of successes  $p_s$  typically of 90% and 100%. Given that expected probability of success of a Bernoulli trial observing  $n_s$  successes out of  $n_\omega$  samples is [52]

$$\mathbb{E}[p_s] = \frac{n_s + 1}{n_\omega + 2} \quad (14)$$

the actual probability of successful navigation  $p_n$  over an edge with  $p_s = 100\%$  successful samples is 95.5%, and similarly  $p_n = 86.3\%$  for thresholds of  $p_s = 90\%$ . Extrapolating over the sequence of edges in a PRM-RL path, the expected success rate is  $p_n^{n_w}$  where  $n_w$  is number of waypoints. In [23], we observe PRM-RL paths with 10.25 waypoints averaged over our three deployment maps, yielding an estimated probability of success of 22.0% for the 90% threshold and 62.3% for the 100% threshold. Therefore, for the lengths of paths we observe in our typical deployment environments, the 100% threshold improves

PRM-RL's theoretical performance to the point that it is more likely to succeed than not, which is what we observe empirically.

#### D. Limitations

While AutoRL can handle moving obstacles [12], PRM roadmaps remain static after construction, causing two failure modes. First, PRM-RL does not replan. If dynamic obstacle avoidance steers a local planner closer to a subsequent waypoint, PRM-RL could replan and provide that waypoint as the next waypoint. Second, large changes in the environment can invalidate edges or create new paths, e.g., adding/removing a wall. Replanning with a roadmap update could handle this scenario, e.g., with iterative reshaping [69].

Another limitation of PRM-RL is that it requires a map. With a sufficiently good local policy, a SLAM algorithm, and an incrementally updatable PRM, it would be possible to build a PRM online by progressively exploring an environment and building the roadmap and SLAM map together. However, while adding online features to a roadmap are certainly feasible, developing an exploration policy is challenging in its own right, and goes hand in hand with improving the quality of the local planner so it can be trusted to execute reliably.

This work focuses on evaluating roadmap construction and performance, so we leave replanning, exploration policies, and online map building for future work.

## VII. CONCLUSION

In this article, we presented PRM-RL, a hierarchical planning method for long-range navigation that combines sampling-based path planning with RL agent as a local planner in very large environments. The core of the method was that roadmap nodes are connected only when the RL agent can connect them consistently in the face of noise and obstacles. This extension of [23] contributed roadmap construction and robot deployment algorithms, along with roadmap connectivity, computational complexity, and navigation performance analysis. We evaluated the method on a differential drive and a car model with inertia used on floormaps from five building, two noisy obstacle maps, and on three physical testbed environments.

We showed that 1) the navigation quality and resilience to noise of the execution policy directly transfers to the hierarchical planner; 2) a 100% success threshold in roadmap construction yields both the highest quality and most computationally efficient roadmaps; and 3) building roadmaps from the noisy SLAM maps that the robot uses at execution time virtually closes the sim2real gap, yielding simulation success rates very similar to those observed on robots. PRM-RL with SLAM maps embed information into the roadmap that the robot uses at execution time, providing a better estimate of performance on the robot. Failure modes included pathologies of the local policy, poorly positioned samples, and sparsely connected roadmaps. In future work, we will examine improved policies with more resistance to noise, better sampling techniques to position the samples strategically, and techniques for improving map coverage with better localization and obstacle maps.

## APPENDIX TABLE OF SYMBOLS

Symbol	Units or Domain	Meaning
$\mathcal{C}$	$\mathbb{R}^{D_c}$	Configuration space of dimension $D_c$
$\mathcal{C}_{\text{free}}$	$\mathbb{R}^{D_c}$	Free portion of configuration space
$\mathcal{S}$	$\mathbb{R}^{D_c+D_t}$	State space of robot plus task state
$\mathcal{T}$	$\mathbb{R}^2 \times \mathbb{S}^1$	Task space for navigation
$\mathcal{T}_{\text{free}}$	$\mathbb{R}^{D_{\mathcal{W}}}$	Free portion of the task space
$\mathcal{W}$	$\mathbb{R}^{D_{\mathcal{W}}}$	Physical workspace of dimension $D_{\mathcal{W}}$
$\mathcal{O}$	$\mathbb{R}^{D_o}$	Observation space of dimension $D_o$
$\mathcal{A}$	$\mathbb{R}^{D_a}$	Action space of dimension $D_a$
$\mathcal{D}$	$S \times A$	Task dynamics $\rightarrow \mathcal{C}$
$\mathcal{N}$	$\mathcal{C} \times A$	Noise model $\rightarrow \mathcal{O}$ or $\mathcal{A}$
$\mathcal{R}$	$O$	Reward model $(\mathcal{G}, r)$
$\mathcal{G}$	$C$	True objective $\rightarrow \mathbb{R}$
$r$	$O$ parametrized with $\theta$	Dense reward $\rightarrow \mathbb{R}$
$r_{\text{name}}$	$O$ parametrized with $\theta_{\text{name}}$	Named reward component $\rightarrow \mathbb{R}$
$\gamma$	$[0..1]$	Discount
$\mathbb{I}$	$\mathbb{B}$	Indicator function $\rightarrow \{1, 0\}$
$L(\mathbf{x})$	$\mathcal{C}$	Task predicate
$F_s(\mathbf{x})$	$\mathcal{C}$	Sensor w/ dyn. $D_s(\mathbf{x})$ & noise $\mathcal{N}_s$
$F_a(\mathbf{x}, \mathbf{a})$	$\mathcal{C} \times A$	Action w/ dyn. $D_a(\mathbf{x})$ & noise $\mathcal{N}_a$
$p_i$	$\mathcal{C}_{\text{free}}$	Waypoint $i$ on path $\mathcal{P}$
$\mathbf{x}_i$	$\mathcal{C}$	Point $i$ along trajectory $\mathcal{T}$
$\mathbf{x}_S$	$\mathcal{C}_{\text{free}}$	Start state
$\mathbf{x}_G$	$\mathcal{C}_{\text{free}}$	Goal state
$d_G$	meters	Goal success distance
$\mathcal{K}_\omega$	$\mathbb{Z}^+$	Max trajectory execution steps
$\mathbf{a}$	$\mathbb{R}^2 \times \mathbb{S}^1$	Diff. drive action $(v, \phi)$ lin, ang vel.
$V_i V_j$	$\mathcal{C}$	Line from $V_i$ to $V_j$ in graph $(V, E)$
$n_\omega$	$\mathbb{Z}^+$	Num. edge connection attempts
$n_s$	$\mathbb{Z}^*$	Num. observed connection successes
$p_s$	$[0..1]$	Edge connection success threshold
$d_\omega$	meters	Max attempted edge connection dist.
$d_\pi$	meters	Policy $\pi$ effective nav. distance
$f_\pi$	$[0..1]$	Policy $\pi$ effective obst. shrinkage
$\rho_\omega$	points/meters <sup>2</sup>	Sampling density per meter
$n$	points	Number of points to sample
$p_n$	$[0..1]$	Probability of successful navigation
$n_w$	$\mathbb{Z}^+$	Number of waypoints on a path
$V_{\mathcal{W}}$	meters <sup>2</sup>	Volume of the workspace
$D_{\mathcal{W}}$	$\mathbb{Z}^+$	Workspace dimensionality
$n_{\mathcal{W}}$	$\mathbb{Z}^+$	Nodes in the workspace
$m_\pi$	$\mathbb{Z}^+$	Neighbors of a node
$n_p$	$\mathbb{Z}^+$	Number of processors
$\sigma_x$	$\mathbb{R}^+$	$\mathcal{N}(0, \sigma_x)$ noise for $x = l, g, v, a$
$\theta_n$	$\mathbb{Z}^+$	Observation trace length

## APPENDIX TRAINING HYPERPARAMETERS

Both actor and critic use the AdamOptimizer with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ,  $\epsilon = 1e-08$ ; the actor's learning rate is  $1e-05$  and the critic's is 0.0005. The actor uses DQDA gradient clipping and the critic uses  $\gamma = 0.995$  with the Huber loss for temporal difference errors. In total, 10 000 initial stabilization steps are followed by a soft target network update of 0.0001 on every step. Our training batch size is 512 and our replay buffer has a capacity of 0.5 million. We train for 5 million steps, but save policies every 25 000 steps and select the best policy over the run.

## ACKNOWLEDGMENT

The authors thank J. Bingham, J. Davidson, B. Ichter, K. Oslund, P. Pastor, O. Ramirez, C. Richards, L. Tapia, A. Toshev, and V. Vanhoucke for helpful discussions and contributions to this project. They also thank the editors and reviewers for their detailed reviews and thoughtful comments.

## REFERENCES

- [1] D. Abel, A. Agarwal, F. Diaz, A. Krishnamurthy, and R. E. Schapire, "Exploratory gradient boosting for reinforcement learning in complex domains," 2016, *arXiv:1603.04119*.
- [2] A. Agha-mohammadi, S. Chakravorty, and N. Amato, "FIRM: Sampling-based feedback motion planning under motion uncertainty and imperfect measurements," *Int. J. Robot. Res.*, vol. 33, no. 2, pp. 268–304, 2014.
- [3] I. Al-Bluwi, T. Siméon, and J. Cortés, "Motion planning algorithms for molecular simulations: A survey," *Comput. Sci. Rev.*, vol. 6, no. 4, pp. 125–143, 2012.

- [4] R. Alterovitz, T. Simeon, and K. Goldberg, "The stochastic motion roadmap: A sampling framework for planning with markov motion uncertainty," in *Proc. Robot.: Sci. Syst.*, Atlanta, GA, USA, Jun. 2007, pp. 246–253.
- [5] N. M. Amato and L. K. Dale, "Probabilistic roadmap methods are embarrassingly parallel," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 1999, vol. 1, pp. 688–694.
- [6] S. Bhatti, A. Desmaison, O. Miksik, N. Nardelli, N. Siddharth, and P. H. Torr, "Playing doom with slam-augmented deep reinforcement learning," 2016, *arXiv:1612.00380*.
- [7] S. Brahmbhatt and J. Hays, "Deepnav: Learning to navigate large cities," 2017, *arXiv:1701.09135*.
- [8] C. Ichniowski and C. Alterovitz, "Parallel sampling-based motion planning with superlinear speedup," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 1206–1212.
- [9] A Look at the U.S. Commercial Building Stock: Results from EIA's 2012 Commercial Buildings Energy Consumption Survey (CBECS). 2012. [Online]. Available: <https://www.eia.gov/consumption/commercial/reports/2012/buildstock/>
- [10] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2722–2730.
- [11] H.-T. Chiang, N. Malone, K. Lesser, M. Oishi, and L. Tapia, "Path-guided artificial potential fields with stochastic reachable sets for motion planning in highly dynamic environments," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2015, pp. 2347–2354.
- [12] H.-T. L. Chiang, A. Faust, M. Fiser, and A. Francis, "Learning navigation behaviors end-to-end with autorl," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 2007–2014, Apr. 2019.
- [13] H. L. Chiang, J. Hsu, M. Fiser, L. Tapia, and A. Faust, "RL-RRT: Kinodynamic motion planning via learning reachability estimators from RL policies," in *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 4298–4305, Oct. 2019.
- [14] B. Donald, P. Xavier, J. Canny, J. Canny, J. Reif, and J. Reif, "Kinodynamic motion planning," *J. ACM*, vol. 40, no. 5, pp. 1048–1066, Nov. 1993.
- [15] A. Dosovitskiy and V. Koltun, "Learning to act by predicting the future," 2016, *arXiv:1611.01779*.
- [16] Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel, "RL<sup>2</sup>: Fast reinforcement learning via slow reinforcement learning," 2016, *arXiv:1611.02779*.
- [17] C. Ekenna, S. A. Jacobs, S. Thomas, and N. M. Amato, "Adaptive neighbor connection for PRMS: A natural fit for heterogeneous environments and parallelism," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 1249–1256.
- [18] M. O'Kelly *et al.*, "F1/10: An open-source autonomous cyber-physical platform," 2019, *arXiv:1611.02779*.
- [19] T. Fan *et al.*, "Getting robots unfrozen and unlost in dense pedestrian crowds," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1178–1185, Apr. 2019.
- [20] A. Faust, H.-T. Chiang, and L. Tapia, "PEARL: Preference appraisal reinforcement learning for motion planning," 2018, *arXiv:1811.12651*.
- [21] A. Faust, N. Malone, and L. Tapia, "Preference-balancing motion planning under stochastic disturbances," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2015, pp. 3555–3562.
- [22] A. Faust, I. Palunko, P. Cruz, R. Fierro, and L. Tapia, "Automated aerial suspended cargo delivery through reinforcement learning," *Artif. Intell.*, vol. 247, pp. 381–398, 2017.
- [23] A. Faust *et al.*, "PRM-RL: long-range robotic navigation tasks by combining reinforcement learning and sampling-based planning," in *Proc. IEEE Int. Conf. Robot. Automat.*, Brisbane, QLD, 2018, pp. 5113–5120.
- [24] D. Fox, W. Burgard, and S. Thrun, "The dynamic window approach to collision avoidance," *IEEE Robot. Autom. Mag.*, vol. 4, no. 1, pp. 23–33, Mar. 1997.
- [25] R. Geraerts and M. H. Overmars, "A comparative study of probabilistic roadmap planners," in *Algorithmic Foundations of Robotics V*. New York, NY, USA: Springer, 2004, pp. 43–57.
- [26] A. Giusti *et al.*, "A machine learning approach to visual perception of forest trails for mobile robots," *IEEE Robot. Autom. Lett.*, vol. 1, no. 2, pp. 661–667, Jul. 2016.
- [27] Gmapping - ROS Wiki, 2019. [Online]. Available: <http://wiki.ros.org/gmapping>
- [28] D. Golovin, B. Solnik, S. Moitra, G. Kochanski, J. Karro, and D. Sculley, "Google vizier: A service for black-box optimization," in *Proc. ACM Int. Conf. Knowl. Discovery Data Mining*, ACM, 2017, pp. 1487–1495.
- [29] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik, "Cognitive mapping and planning for visual navigation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Honolulu, HI, 2017, pp. 7272–7281.
- [30] K. Hauser, T. Bretl, J. Latombe, and B. Wilcox, "Motion planning for a sixlegged lunar robot," in *Proc. 7th Int. Workshop the Algorithmic Found. Robot.*, 2006, pp. 16–18.
- [31] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Phys. Rev. E*, vol. 51 no. 5, 1995, Art. no. 4282.
- [32] D. Hsu, R. Kindel, J.-C. Latombe, and S. M. Rock, "Randomized kinodynamic motion planning with moving obstacles," *Int. J. Robot. Res.*, vol. 21, no. 3, pp. 233–256, 2002.
- [33] D. Hsu, J.-C. Latombe, and H. Kurniawati, "On the probabilistic foundations of probabilistic roadmap planning," in *Robotics Research*. New York, NY, USA: Springer, 2007, pp. 83–97.
- [34] A. Irpan, "Deep reinforcement learning doesn't work yet," 2018. [Online]. Available: <https://www.alexirpan.com/2018/02/14/rl-hard.html>
- [35] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *J. Artif. Intell. Res.*, vol. 4, pp. 237–285, 1996.
- [36] S. Karaman and E. Frazzoli, "Sampling-based algorithms for optimal motion planning," *Int. J. Robot. Res.*, vol. 30, no. 7, pp. 846–894, 2011.
- [37] Y. Kato, K. Kamiyama, and K. Morioka, "Autonomous robot navigation system with learning based on deep q-network and topological maps," in *Proc. IEEE/SICE Int. Symp. Syst. Integration*, Dec. 2017, pp. 1040–1046.
- [38] L. E. Kavraki and J. C. Latombe, "Probabilistic roadmaps for robot path planning," in *Practical motion planning in robotics: current approaches and future challenges*, K. Gupta and A. P. Pobil, Eds. Hoboken, NJ, USA: Wiley, 1998, pp. 33–53.
- [39] L. E. Kavraki, P. Švestka, J. C. Latombe, and M. H. Overmars, "Probabilistic roadmaps for path planning in high-dimensional configuration spaces," *IEEE Trans. Robot. Autom.*, vol. 12, no. 4, pp. 566–580, Aug. 1996.
- [40] O. Khatib, "Real-time obstacle avoidance for manipulators and mobile robots," *Int. J. Robot. Res.*, vol. 5, no. 1, pp. 90–98, 1986.
- [41] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1238–1274, 2013.
- [42] J. Kuffner and S. LaValle, "RRT-connect: An efficient approach to single-query path planning," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2000, vol. 2, pp. 995–1001.
- [43] S. M. LaValle, *Planning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2006.
- [44] S. M. LaValle and J. J. Kuffner, "Rapidly-exploring random trees: Progress and prospects," in *Algorithmic and Computational Robotics: New Directions*. Boston, MA, USA: A. K. Peters, 2000, pp. 293–308.
- [45] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *J. Mach. Learn. Res.*, vol. 17, no. 39, pp. 1–40, 2016.
- [46] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," in *Proc. Int. Conf. Learn. Representations*, 2016. [Online]. Available: <https://deepmind.com/research/publications/continuous-control-deep-reinforcement-learning>
- [47] P. Long, T. Fan, X. Liao, W. Liu, H. Zhang, and J. Pan, "Towards optimally decentralized multi-robot collision avoidance via deep reinforcement learning," in *Proc. IEEE Int. Conf. Robot. Autom.*, Brisbane, Australia, May 21–25, 2018, pp. 6252–6259.
- [48] E. S. Low, P. Ong, and K. C. Cheah, "Solving the optimal path planning of a mobile robot using improved q-learning," *Robot. Auton. Syst.*, vol. 115, pp. 143–161, 2019.
- [49] N. Malone, A. Faust, B. Rohrer, R. Lumia, J. Wood, and L. Tapia, "Efficient motion-based task learning for a serial link manipulator," *Trans. Control Mech. Syst.*, vol. 3, no. 1, pp. 25–35, 2014.
- [50] N. Malone, K. Manavi, J. Wood, and L. Tapia, "Construction and use of roadmaps that incorporate workspace modeling errors," in *Proc. IEEE Int. Conf. Intell. Robot. Syst.*, Nov. 2013, pp. 1264–1271.
- [51] N. Malone, B. Rohrer, L. Tapia, R. Lumia, and J. Wood, "Implementation of an embodied general reinforcement learner on a serial link manipulator," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 862–869.
- [52] N. D. Megill and M. Pavicic, "Estimating bernoulli trial probability from a small sample," 2011, *arXiv:1105.1486*.
- [53] M. Wise, M. Ferguson, D. King, E. Diehr, and D. Dymesich, "Fetch & freight: Standard platforms for service robot applications," in *Proc. Workshop Auton. Mobile Service Robots Held Int. Joint Conf. Artif. Intell.*, 2016. [Online]. Available: <https://fetchrobotics.com/wp-content/uploads/2018/04/Fetch-and-Freight-Workshop-Paper.pdf>
- [54] P. Mirowski *et al.*, "Learning to navigate in complex environments," 2016, *arXiv:1611.03673*.
- [55] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [56] A. Wahid, A. Toshev, M. Fiser, and T. E. Lee, "Long range neural navigation policies for the real world," *Intell. Robots Syst. 2019 IEEE/RSJ Int. Conf.*, pp. 82–89, 2019.
- [57] K. Mülling, J. Kober, and J. Peters, "A biomimetic approach to robot table tennis," *Adaptive Behav.*, vol. 19, no. 5, pp. 359–376, 2011.

- [58] B. Paden, M. Cáp, S. Z. Yong, D. S. Yershov, and E. Frazzoli, "A survey of motion planning and control techniques for self-driving urban vehicles," in *IEEE Trans. Intell. Vehicles*, vol. 1, no. 1, pp. 33–55, Mar. 2016.
- [59] J.-J. Park, J.-H. Kim, and J.-B. Song, "Path planning for a robot manipulator based on probabilistic roadmap and reinforcement learning," *Int. J. Control. Autom. Syst.*, vol. 5, pp. 674–680, 2008.
- [60] M. Pfeiffer, M. Schaeuble, J. I. Nieto, R. Siegwart, and C. Cadena, "From perception to decision: A data-driven approach to end-to-end motion planning for autonomous ground robots," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2017, pp. 1527–1533.
- [61] S. Rodríguez, J.-M. Lien, and N. M. Amato, "A framework for planning motion in environments with moving obstacles," in *Proc. IEEE Int. Conf. Intel. Rob. Syst.*, 2007, pp. 3309–3314.
- [62] A. Seff and J. Xiao, "Learning from maps: Visual common sense for autonomous driving," 2016, *arXiv:1611.08583*.
- [63] B. Siciliano and O. Khatib, *Springer Handbook of Robotics*. New York, NY, USA: Springer, 2016.
- [64] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.
- [65] L. Tai, G. Paolo, and M. Liu, "Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2017, pp. 31–36.
- [66] L. Tapia, S. Thomas, and N. M. Amato, "A motion planning approach to studying molecular motions," *Commun. Inf. Syst.*, vol. 10, no. 1, pp. 53–68, 2010.
- [67] J. van den Berg, P. Abbeel, and K. Goldberg, "Lqg-mp: Optimized path planning for robots with motion uncertainty and imperfect state information," *Int. J. Robot. Res.*, vol. 30, no. 7, pp. 895–913, 2011.
- [68] A. Yahya, A. Li, M. Kalakrishnan, Y. Chebotar, and S. Levine, "Collective robot reinforcement learning with distributed asynchronous guided policy search," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Vancouver, BC, 2017, pp. 79–86.
- [69] E. Yoshida, C. Esteves, I. Belousov, J. Laumond, T. Sakaguchi, and K. Yokoi, "Planning 3-d collision-free dynamic robotic motion through iterative reshaping," *IEEE Trans. Robot.*, vol. 24, no. 5, pp. 1186–1198, Oct. 2008.
- [70] J. Zhang, J. T. Springenberg, J. Boedecker, and W. Burgard, "Deep reinforcement learning with successor features for navigation across similar environments," in *Proc. IEEE Int. Conf. Intel. Robot. Syst.*, 2017, pp. 2371–2378.
- [71] Y. Zhu *et al.*, "Target-driven visual navigation in indoor scenes using deep reinforcement learning," 2016, *arXiv:1609.05143*.
- [72] Y. Zhu *et al.*, "Target-driven visual navigation in indoor scenes using deep reinforcement learning," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2017, pp. 3357–3364.



**Anthony Francis** (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the Georgia Institute of Technology, Atlanta, GA, USA, in 1991, 1996, and 2000, respectively, all in computer science, along with a Certificate in Cognitive Science in 1999.

He is currently a Senior Software Engineer with Robotics at Google, Mountain View, CA, USA, specializing in reinforcement learning for robot navigation. Previously, he worked on emotional long-term memory for robot pets at Georgia Tech's PEPE robot pet project, on models of human memory for information retrieval at Enkia Corporation, and on large-scale metadata search and 3-D object visualization at Google.

Dr. Francis won the ICRA 2018 Best Paper Award for Service Robotics. His work has been featured in the New York Times.



**Aleksandra Faust** (Senior Member, IEEE) received the master's degree in computer science from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2004, and the Ph.D. degree (with distinction) in computer science from the University of New Mexico, Albuquerque, NM, USA, in 2014.

She is currently a Staff Research Scientist with Robotics, Google, Mountain View, CA, USA, specializing in robot motion planning and reinforcement learning.

Dr. Faust's work has been featured in the New York Times, ZdNet, and was awarded the Best Paper in Service Robotics at ICRA 2018.



**Hao-Tien (Lewis) Chiang** (Member, IEEE) received his M.S. degree in physics and Ph.D. in computer science from the University of New Mexico, Albuquerque, NM, USA, in 2015 and 2020, respectively.

He was also a Student Researcher at Robotics at Google from 2018 to 2019. His research focuses on improving and integrating robot motion planning and machine learning. His work combines efficient search techniques from motion planning with noise-tolerant, adaptive reinforcement learning. This line of research resulted in his work being featured in the Google AI Blog, IEEE Spectrum PC Magazine and VentureBeat.com. He co-organized the popular Third Workshop in Machine Learning in the Planning and Control of Robot Motion at ICRA 2018 and two Becoming A Robot Guru workshops at RSS 2016 and WAFR 2018.



**Jasmine Hsu** (Member, IEEE) received the B.A. degree in cognitive science from the University of Virginia, Charlottesville, VA, USA, in 2012, and the M.S. degree in computer science from New York University, New York City, NY, USA, in 2015.

She previously worked in the defense industry and has been a Software Engineer with Robotics at Google, Mountain View, CA, USA, since 2016. Her previous work has been focused on reinforcement learning for grasping, learning representations, and currently motion-planning.



**J. Chase Kew** (Member, IEEE) received the B.S. degree in computer science and mechanical engineering from the California Institute of Technology, Pasadena, CA, USA, in 2017.

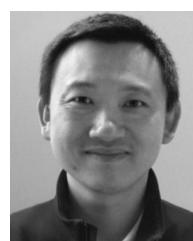
She is currently a Software Engineer with Robotics, Google, Mountain View, CA, USA, working on machine learning for robotic navigation.



**Marek Fiser** received the master's degree in computer science from Purdue University, West Lafayette, IN, USA, in 2015.

He is currently a Software Engineer with Robotics at Google, Mountain View, CA, USA, working on systems that can learn navigation policies with reinforcement learning, including creation and integration of simulated environments, designing and training agents using RL, and deploying learnt policies on real robots.

Mr. Fiser's work was awarded the Best Paper in Service Robotics in 2018.



**Tsang-Wei Edward Lee** received the B.S. degree in computer science and electrical engineering from the University of California, Riverside, Riverside, CA, USA, in 2002.

He is currently a Test Engineer with Robotics at Google, Mountain View, CA, USA, designing test plans, supporting robot operations, and conducting on robot experiments.